63-3-1

Department of Psychology
University of Illinois
Urbana, Illinois

FINAL REPORT ON CONTEXT EFFECTS IN PSYCHOPHYSICAL JUDGMENTS
(INTRODUCTION and PART I: THE UP-AND-DOWN METHOD)

William E. Kappauf

Contract No. DA-49-007-MD-877, concerning,
Context Effects in Psychophysical Judgments,

supported in part by

The Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

William E. Kappauf
Principal Investigator

Department of Psychology
University of Illinois
Urbana, Illinois

FINAL REPORT ON CONTEXT EFFECTS IN PSYCHOPHYSICAL JUDGMENTS
(INTRODUCTION and PART I: THE UP-AND-DOWN METHOD)

William E. Kappauf

Contract No. DA-49-007-MD-877, concerning,
Context Effects in Psychophysical Judgments,
supported in part by
The Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

William E. Kappauf
Principal Investigator

TABLE OF CONTENTS

INTRODUCTION

PART I: THE UP-AND-DOWN METHOD

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

AD NO. ____________________ ACCESSION NO. ____________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-007-MD-877

6. Supported by: Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

ABSTRACT

AD NO. ____________________ ACCESSION NO. ____________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-007-MD-877

6. Supported by: Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

A B S T R A C T

AD NO. ____________________ ACCESSION NO. ____________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-007-MD-877

6. Supported by: Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

ABSTRACT

AD NO. ________________ ACCESSION NO. ________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-007-MD-877

6. Supported by: Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

A B S T R A C T

AD NO. ____________________ ACCESSION NO. ____________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-007-MD-877

6. Supported by: Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

ABSTRACT

AD NO. ________________ ACCESSION NO. ________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-007-MD-877

6. Supported by: Research and Development Division
Office of the Surgeon General
Department of the Army
Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

ABSTRACT

AD NO. ____________________ ACCESSION NO. ____________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-C07-MD-877

6. Supported by: Research and Development Division
   Office of the Surgeon General
   Department of the Army
   Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

ABSTRACT

AD NO. ____________________ ACCESSION NO. ____________________

1. Preparing Institution: University of Illinois, Department of Psychology

2. Title of Report: Final Report on Context Effects in Psychophysical Judgments. (Introduction and Part I: The Up-and-Down Method). by William E. Kappauf.

3. Principal Investigator: William E. Kappauf

4. Pages 1 to 90 8 illustrations 15 March 1963

5. Contract Number: DA-49-007-MD-877

6. Supported by: Research and Development Division
   Office of the Surgeon General
   Department of the Army
   Washington 25, D. C.

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, considerable attention was given to the method as such.

The first--and present--part of our report discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. Particular topics include: (a) origin and general description of the up-and-down method; (b) an illustration of its use in the study of differential judgment; (c) the "expected" distribution of up-and-down trials and the influence of stimulus step size upon this distribution; (d) estimation of σ and μ of the probability of response function; (e) selection of stimulus step size; (f) use of the up-and-down method in situations where each subject participates for only one trial; (g) use of several concurrent up-and-down series for extended testing of a single subject; (h) empirical data on the reliability of estimates of μ, under uses cited in both (f) and (g); (i) an experimental comparison of the method of constant stimuli and the up-and-down method in the measurement of judgment bias; (j) some points of similarity and difference between the method of limits and the up-and-down method; (k) potential scope and limitations of the up-and-down method in psychological research.

The second part of our report will discuss our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

NOTE: Copies of this report are filed with the Armed Services Technical Information Agency, Arlington Hall Station, Arlington 12, Virginia, and may be obtained from that agency by qualified investigators working under Government contract.

INTRODUCTION

Research conducted under this contract has concerned context effects in differential judgments. The principal method employed has been the Up-and-Down Method. A starting premise was that this method would have distinct advantages over the traditional method of constant stimuli for the study of any kind of judgment bias, and hence for the study of context effects in particular. Because this work represented an initial application of the up-and-down method to the study of differential discrimination, it was desirable that we devote some attention to the method as such. Our report therefore falls into two main parts. The first of these discusses the up-and-down method, various ways in which we employed it, our specific attempts to improve its usefulness in discrimination studies, and our comparative tests of this method and the method of constant stimuli. The second part of the report discusses our experimental results on the context problem: our studies of context effects in the multiple standards situation, and our examination of the effects which individual trials or sequences of previous trials have upon a current discrimination.

1. THE DIFFERENTIAL JUDGMENT SITUATION.

Our basic test situation has been one in which the subject is required to compare two stimuli and to indicate whether the second is heavier or lighter than the first, louder or less loud, longer or shorter, nearer or farther away. In all cases, the subject has been limited to two oppesed judgments, as just illustrated. Judgments of "equal" were never allowed.

Throughout the discussions which follow, we refer to the function which relates the probability of either of these alternative responses to the stimulus dimension under study as the "psychometric function" or the "probability of response curve". We designate the first stimulus in the pair as the "standard" and the second as the "comparison". We use the terms "point of objective equality" (POE) and "point of subjective equality" (PSE) in the conventional way, POE referring to that comparison stimulus which is objectively equal to the standard, and PSE referring to that comparison stimulus which the subject judges to be the equal of the standard. Judgment bias is given by the quantity (PSE-POE).

2. CONTEXT, AND CONTEXT EFFECTS.

Every judgment or response to particular stimuli is made, we shall say, in the context of preceding and concurrent stimuli and in the context of preceding and concurrent responses. This is to assert, for the typical laboratory, psychophysical experiment, that the judgment on any individual trial is made within the context of previous trials and the testing situation.

We are going to use the term "context effect" throughout this report to apply to any bias of a present judgment which is a function of previous trials.

Other authors have at times used expressions such as "series effect" (e.g., Woodworth and Schlosberg, 1954),"central tendency effect" (e.g. Hollingworth, 1910), negatively directed constant error function (Koester and Schoenfeld, 1946), or "response dependencies" (e.g. Senders and Sowards, (1952), to refer to portions, aspects or features of what we here call "context effect". Our preference for "context effect" stems from our interest in having a general term to apply to any and all effects of prior trials on a current judgment. We thus bring together and include under this term effects which have not necessarily been considered together in the literature. We include effects which are session-long or long-range, in that they are associated with the entire collection of prior trials, as well as effects which are very short-range in that they are associated with the immediately preceding trial or at most a few preceding trials.

3. STIMULUS CONTEXT AND JUDGMENT CONTEXT.

We find it convenient in some discussions to distinguish between stimulus context on the one hand, and judgment or response context on the other. Stimulus context consists of the stimuli presented over some prescribed set of prior trials, whatever set happens to be of interest in the study. Judgment context consists similarly of those responses made over some set of prior trials of interest in the study. Although stimulus context and judgment context typically cannot be manipulated independently of each other, the experimenter may so design his experiment that he controls one or the other specifically, or in some special cases controls them jointly.

The method of constant stimuli, and any other psychophysical method which employs a pre-arranged stimulus series permits the direct control of stimulus context. Session-long stimulus context is controlled and can be manipulated in terms of the session-long distribution of comparison stimuli. In most experiments by the method of constant stimuli, this distribution of comparison stimuli is rectangular and centered about the POE, but the effect of altering this distribution, i.e. of altering session-long stimulus context, has been investigated (e.g. Harris, 1948). Short-term stimulus context, on the other hand, may be varied and studied by organizing the pre-arranged stimulus series in such a way that particular comparison stimuli are used on trials preceding those of interest. (See Fernberger, 1920; Koester and Schoenfeld, 1946).

Experiments dealing with short-term judgment context have made a relatively recent appearance in the literature, typically under titles referring to the non-independence of successive responses (e.g. Verplanck _et al_, 1952). Problems concerned with long-term or session-long judgment context on the other hand have attracted little if any interest. Perhaps this will change, however, now that we have a formal way of controlling session-long judgment context through the up-and-down method. The sequential nature of this method makes it possible to regulate or control the session-wide distribution of judgments which the subject makes (in a two-alternatives response situation). Normal application of the method assures that, in the long run, the subject will use an equal number of "greater than" and "less than" judgments. Modifications of the method provide situations in which the ratio of the number of "greater than" judgments to "less than" judgments will be 2:1, 3:1, etc.

In the foregoing terms, the method of constant stimuli is a method in which session-wide stimulus context is controlled by the experimenter, and the judgment distribution or judgment context is a function of the subject's responses to those stimuli. The up-and-down method, on the other hand, is a method in which session-wide judgment context is controlled in terms of the sequential program which the experimenter adopts for the study, and the stimulus distribution or stimulus context is a function of the subject's discriminative behavior. To compare the method of constant stimuli and the up-and-down method, then, is to a compare the effect of controlling stimulus context on the one hand and of controlling judgment context on the other.

4. ORDER OF TOPICS AND EXPERIMENTS IN THIS REPORT.

The work of this project extended from July 1957 to August 1961. Related work using the up-and-down method had been undertaken in our laboratory as early as 1952. The report which follows includes a discussion of some of these "pre-project" studies, when they are relevant, as well as the project research.

The report is organized not chronologically, but by topics or problems. This means that early, exploratory studies with smaller groups of subjects are occasionally interspersed with later, more thorough studies. It means also that our entire discussion of the up-and-down method precedes the research portion of the report, even though not all of our context studies benefited from planning based upon all that we now know about the method. We believe, however, that the chosen organization will be found the more useful.

References

1. Fernberger, S. W. Interdependence of judgments within the series for the method of constant stimuli. J. exper. Psychol., 1920, 3, 126-150.
2. Harris, J. D. Discrimination of pitch: suggestions toward method and procedure. Amer. J. Psychol., 1948, 61, 309-322
3. Hollingworth, H. L. The central tendency of judgment. J. philos. psychol. sci. Meth., 1910, 7, 461-469.
4. Koester, T. and Schoenfeld, W. N. The effect of context upon judgments of pitch differences. J. exper. Psychol., 1946, 36, 417-430.
5. Senders, Virginia L., and Sowards, A. Analysis of response sequences in the setting of a psychophysical experiment. Amer. J. Psychol., 1952, 65, 358-374
6. Verplanck, W. S., Collier, G. H. and Cotton, J. W. Non-independence of successive responses in measurements of the visual threshold. J. exper. Psychol., 1952, 44, 273-282.
7. Woodworth, R. S., and Schlosberg, H. Experimental Psychology, Rev. Ed. New York: Holt, 1954. pp xi + 948

PART I: THE UP-AND-DOWN METHOD

1. ORIGINS AND GENERAL DESCRIPTION OF THE UP-AND-DOWN METHOD

The Up-and-Down Method was developed by non-psychologists, but early discussions of the use of the procedure recognized its potential application to a wide range of measurement problems, including those of psychophysics (Anon., 1944; McCarthy, 1947). The up-and-down method is a general measurement method, and like other measurement methods with which the psychologist is familier, it is both a routine for collecting observations and a numerical or statistical procedure for deriving desired measurements from the data. In this particular case, the method of data collection was introduced by members of the Explosives Research Laboratory at Bruceton, Pennsylvania, while the associated statistical procedures were developed by the Statistical Research Group of Princeton University (Anon., 1944; Anderson, McCarthy and Tukey, 1946). The unique feature of the method is the special sequential character of its data collecting operation. A consequent and highly desirable property of the method is its efficiency: relatively few trials or observations are required per measurement.

The up-and-down method applies in situations where two alternative responses, X and Y, are possible. These alternative responses may be of the sort "detection" vs. "non-detection", "second weight lighter than the first" vs. "second weight heavier than the first", etc. We arrange suitable stimulus conditions for the first trial and conduct that trial. If response X occurs, then according to the up-and-down method, our next trial is run under stimulus conditions which, by one step on our stimulus scale, are more favorable to the occurrence of response Y. If, on the other hand, response Y occurs, our next trial is run under stimulus conditions which, by one step on the stimulus scale, are more favorable to the occurrence of response X. Throughout a complete series of trials, each successive trial is scheduled on the basis of these same rules. This programming of trials is said to be "sequential" because the stimulus conditions used on each new trial are contingent on the outcome of the previous trial.

The up-and-down routine just outlined presumes that the probability of response X increases monotonically from 0 to 1.00 (and that the probability of response Y decreases from 1.00 to 0), as successive steps are taken along the stimulus scale. When the psychometric function is of this character, then the up-and-down method assures that the series of trials will see-saw up and down the stimulus scale, and that most of the trials will be conducted in the vicinity of that stimulus level where the probability of response X and the probability of response Y are equal. Very few trials are conducted at extreme stimulus levels, because whenever the trial series moves to such a level, the response which is there highly probable drives the series back to less extreme levels.

When the series of observations has been terminated, application of the numerical procedures developed by the Princeton Research Group provides an estimate of the stimulus level which marks the 50% point, or point of transition from response X to response Y. Further calculations lead to an estimate of the variability of behavior in the region of this transition. These estimates are based on two assumptions: that the psychometric function is a cumulative normal, and that the experimenter has chosen equally spaced stimulus levels on the scale which provides that normality.

It will have been apparent from this description that the up-and-down method uses discrete stimulus values. The stimuli on any given trial are fixed and unchanging during that trial. On this basis the method may be classed as one of the "constant stimulus" methods. It may also be classified as one of the "frequency methods" in that measurements obtained by the method are computed from the observed frequency of occurrence of the alternative responses under each stimulus condition. By the Princeton Research Group it was identified as a member of the class of "staircase methods", i.e. methods of a sequential sort which include among others, the method of limits (See Anderson et al, 1946).

2. SOME SIMILAR PROCEDURES.

It was not long after the development of the Bruceton-Princeton procedures that an up-and-down scheme for stimulus control in a psychophysical situation was independently devised and introduced by Bekesy (1947) and by Oldfield (1949).

Bekesy incorporated the scheme in his "new audiometer". In this audiometer, the intensity of the signal for which the patient was listening was increased by 2 db. steps every .86 seconds as long as the patient kept a response key depressed indicating that he heard nothing. The signal intensity was decreased by similar steps during all periods when the key was released indicating that the patient could hear the signal. All up and down shifts of signal intensity from levels just above threshold to levels just below threshold were recorded graphically. Then as the audiometer signal was caused to sweep the auditory frequency range from one end to the other, the recorder traced out a saw-toothed but easily interpreted picture of the patient's audiogram. Because of its great convenience, this audiometer has found widespread use in clinical and in research work. Newer forms of it permit the use of rates of intensity change up to 5 db. per second in steps of .25 db.

Oldfield's concern was with measurements of the absolute visual threshold. He devised a continuous, motor-driven intensity control which altered the intensity of the visual stimulus at the rate of one-half log unit per second. The subject kept a response button depressed as long as the stimulus was visible, thereby causing the motor to decrease the stimulus intensity. When the stimulus was no longer visible, the subject released the button. This reversed the motor which then gradually returned the stimulus intensity to higher levels. Again, as in Bekesy's audiometer, the subject's response kept the stimulus in the vicinity of the threshold.

Thus Bekesy and Oldfield hit upon methods of data collection which were conceptually similar to the testing procedure of the Bruceton group. Stimulus intensity at any moment was a function of the subject's response to the stimulus intensity which had prevailed a moment before, and the stimulus control "hunted" in the vicinity of that stimulus level where the probability of detection was .50.

Three rather important differences in method and objectives are to be noted, however, between the work at Bruceton and that in the Bekesy and Oldfield laboratories:

(1) The Bruceton procedure involved a series of discrete trials, whereas the task for the observers in the Bekesy and Oldfield situations was a continuous observing task.

(2) The object of the Bruceton procedure was to get a reasonable estimate of the parameters of the probability of response function at minimum cost, that is within a minimum number of trials and within a very limited number of up-and-down crossings of the 50% point. This required a formal, numerical analysis of the data. For Bekesy and Oldfield, on the

other hand, the rate of change of stimulus intensity was sufficiently rapid and the intended observation period sufficiently long that the stimulus series would clearly cross the 50% point a great many times. With many stimulus reversals to mark the threshold, no elaborate or specific method of threshold computation was necessary. In fact, for many purposes the graphic record has often been sufficient in itself.

(3) The interest of the Bruceton group was in measurement with regard to some fixed probability of response function, while Bekesy and Oldfield were concerned with extending their observations long enough in time to observe changes in the 50% point as a function of some variable--sound frequency in Bekesy's case, observing time in Oldfield's case. It is for this reason that the Bekesy-Oldfield procedure is frequently referred to as a "tracking method". It permits one to follow or track changes in the value of a sensory parameter. Of interest is the fact that an historically earlier use of the same basic procedure in the area of medicine also had a tracking emphasis--the tracking of blood pressure changes under various conditions of activity on the part of the subject or patient (See Lange, 1943). By contrast, the computational procedures of the up-and-down method assume that all observations have been drawn from the same unchanging population--that one set of response probabilities applies for the entire record being analyzed.

Psychological research over the past 15 years has seen increasingly frequent adaptations of the Bekesy-Oldfield type of procedure. Both stepwise (e.g. Gourevitch *et al*, 1960) and continuous (e.g. Blough, 1956, 1957, Evans, 1961) stimulus control have been used. Unequal stepping in the two directions has been employed (e.g. Koh and Teitelbaum, 1961) as well as the more usual equal stepping. When it has been desirable to quantify the graphic records, suitable methods for computing thresholds have been devised by individual authors (e.g. Blough, 1957; Loeb and Dickson, 1961; Koh and Teitelbaum, 1961). Clearly the Bekesy-Oldfield technique is already well established as a most efficient way to collect data in a wide variety of test situations, the not too cogent concerns of Brown and Cane (1959) notwithstanding.

Under some of the foregoing variations, the Bekesy-Oldfield technique merges with the up-and-down method at the level of the programming of trials, but we shall maintain distinctions here by identifying the up-and-down method with its specific computational routines. So considered, we recognize that the up-and-down method has been used and examined relatively little by psychologists. Probably the primary reason for this is the fact that until recently (Cornsweet, 1962) the only generally available discussion of the method to appear is that given in the statistics text by Dixon and Massey (1951, 1957).

Let us turn then to an account of some of the important features and properties of the up-and-down method.

3. ADVANTAGES SEEN FOR THE UP-AND-DOWN METHOD.

Among the advantages which are seen for the up-and-down method over other psychophysical methods of the "constant" group, we may list the following.

(1) Flexibility and simplified experimental planning: the up-and-down method assures that the series of test trials will itself "hunt" for that stimulus level which marks the transition from response X to response Y, so that pilot studies to locate this level approximately are unnecessary.

(2) Efficiency: measurements of any given reliability should be obtained in fewer trials by the up-and-down method than by other methods (see Brownlee, _et al_, 1953).

(3) Opportunity to determine several limens concurrently: because of the need for fewer trials per measurement, several concurrent measures may be taken in the same experimental session through the use of separate, concurrent up-and-down series (see especially Part II of this report).

(4) Simplified automatic programming: successive trials for a single up-and-down series (or for each of several concurrent series) may be programmed through the use of an add-and-subtract stepper (see Appendix).

One further advantage which may be anticipated, but which requires demonstration, is that the up-and-down method should be relatively free from the kind of bias which arises in the method of constant stimuli from experimenter-determined stimulus context (see page 3 above). It was indeed this particular bias question which we first explored and which led us into our general study of context effects employing the up-and-down method.

Because all of our studies have dealt with some aspect of the problem of bias in differential judgment, the following discussion of details of the up-and-down method is concerned with the manner in which the method may be adapted most effectively to the study of differential discrimination. Extensions of the discussion to other applications should be obvious and therefore, with only a few exceptions, will be left without comment.

4. ILLUSTRATIVE APPLICATION OF THE UP-AND-DOWN METHOD IN THE COLLECTION OF DATA ON DIFFERENTIAL JUDGMENT.

Let us suppose a test situation in which a subject hears two tones in temporal succession. He must report whether the second tone, the comparison, is louder or softer than the first tone, the standard. For successive trials in the up-and-down series, the standard is always the same in intensity, say 90 db. The comparison, however, may be at any one of a number of prearranged intensity levels. One of these levels is chosen for the first trial, say 94 db. If the subject reports the comparison on that trial to be louder than the standard, the experimenter conducts the second trial with a comparison which is one step, say 2 db., below that used on the first trial. If the subject reports the first comparison to be softer or less loud than the standard, the comparison used on the second trial is one step more intense than that used on the first trial. In a similar way the comparison level for every subsequent trial is dependent upon the subject's report on the just previous trial.

A series of 25 differential judgment trials conducted according to this routine is shown in Figure 1. Successive trials, are numbered from left to right across the figure. Intensity levels of the comparison stimulus are shown at the left. The standard was 90 db. On the first trial, the comparison stimulus was 94 db. and it was judged louder than the standard, so an "L" is shown for trial 1 opposite 94 db. The second trial, using a comparison of 92 db., also led to a judgment of louder, so the third trial was conducted using a comparison of 90 db. The first judgment of softer came when 88 db was used as the comparison, so far the fifth trial the comparison was raised back to 90 db. The remaining trials in the series continued in this same manner and involved comparison intensities which ranged between 84 and 90 db.

Clearly over any series, such as that shown in the figure, the comparison stimulus intensity, being contingent upon the subject's responses, will shift up and down irregularly, but will never wander too far away from that comparison level which appears equal to the standard in loudness. By the very nature of the sequential character of the stimulus series, trials are concentrated at those stimulus levels where the subject is shifting from judgments of "louder" to judgments of "softer". For the present data, this point of shift is at about 87 db.

Intensity level of the comparison stimulus:	Trials																									Total distribution of trials.	Distribution of	
	1				5					10					15					20					25		"Louder" responses	"Softer" responses
96																												
94	L																									1	1	
92		L																								1	1	
90			L		L												L		L							4	4	
88				S		L		L		L				L		S		S		L		L		L		10	7	3
86							S		S		L		S		S						S		S		S	8	1	7
84												S														1		1
82																												
80																												

Note that except for early L trials these two distributions would be alike.

Figure 1: Illustrative record showing application of the up-and-down method in a loudness discrimination situation. The standard for the series is a tone of 90db. Comparison tones available differ in steps of 2db. The record shows the outcome of each trial over a series of 25 trials. L means that the subject reported the comparison to be louder than the standard, S that he reported it to be softer.

To the right in the figure, three summary frequency distributions are shown. The first indicates the frequency with which trials were conducted at each of the comparison stimulus levels. The second indicates the frequency distribution of trials on which the subject reported "louder". These are the reports which drove the up-and-down series down for subsequent trials. In the third column is shown the frequency distribution of the trials on which the subject reported "softer". These reports drove the series up for all subsequent trials. Note that the frequency distributions in the last two columns are very much alike: in fact they are only different to the extent that the series does not end with a response which would put the next trial back at the very same level where the series started. How similar these distributions are depends in part on whether the experimenter happens to start the series at a level which is close to the subject's point of subjective equality (PSE). In the present case it is clear that the series started above the PSE, so there were more louder judgments in the full series than there were softer judgments.

As it turns out, a feature of this method which it is important to recognize is that the smaller the size of stimulus step used in the comparison series, the closer the trials are concentrated in the vicinity of the 50-50 point. In order to observe this, it is first necessary for us to examine the manner in which the expected distribution of trials may be calculated from a known psychometric function.

5. THE EXPECTED DISTRIBUTION OF UP-AND-DOWN TRIALS AS DERIVED FROM RESPONSE PROBABILITIES AT EACH TEST LEVEL.

When trials are conducted at successive levels up and down the stimulus scale as a function of the subjects's responses, the frequency distribution of trials over those levels is a direct function of the subject's probability of judging "louder" and "less loud" at each level. Consider the following example:

An experimenter chooses four stimulus levels at which the probabilities that the subject will respond "louder" and "softer"are those given in Table 1. For every 100 trials conducted at stimulus level 3, the expected number of "louder" judgments is 90 and of "softer" judgments is 10, as shown in columns 4 and 5 of the table. The 10 judgments of "softer" each precede, under the up-and-down routine, trials which are

Table 1

COMPUTATION OF THE EXPECTED DISTRIBUTION OF TRIALS IN AN UP-AND-DOWN SERIES.

Stimulus Level	Response Probabilities		Expected response frequencies given 100 trials at level #3		Expected frequency distribution of trials: sum of two previous cols.
	for "louder" response	for "softer" response	"louder" responses	"softer" responses	
#4	1.00	.00	10		10
#3	.90	.10	90	10	100
#2	.25	.75	30	90	120
#1	.00	1.00		30	30

(For discussion, see text)

conducted at level 4 But there the probability of a judgment of "louder" is 1.00, and so these 10 trials are all followed by trials which are back again at stimulus level 3. Clearly if 100 trials are to be conducted altogether at level 3, and 10 are preceded by trials at level 4, the other 90 must be preceded by "softer" responses at level 2. Hence the expected number of "softer" responses at level 2 is 90. These 90 responses, however, constitute .75 of the expected total number of trials at level number 2. So the latter must be 120. In other words, the expected number of "louder" judgments at level 2 would be 30. Each of these 30 "louder" judgments at level 2 leads the experimenter to conduct a new trial at level 1. All of these 30 trials at level 1 must involve "softer" judgments, because at this level the probability of "softer" is given as 1.00.

Thus out of a total of 260 trials, conducted at the levels shown and with the response probabilities indicated, the number of trials expected at each of the stimulus levels, 4, 3, 2, and 1, is respectively 10, 100, 120 and 30.

This serially developed calculation may be applied to any set of probabilities and to cases with any number of stimulus levels. Hence, it is a simple matter to determine the expected distribution of trials for any up-and-down test, given the psychcmetric function and the intended test levels.

To those familiar with other developments in measurement theory, it will come as no surprise that the model used in the development of estimation formulae for the up-and-down method specifies that the probability of response curve be a cumulative normal. It further specifies that the stimulus levels used for testing be equally spaced on the stimulus scale, whatever that scale be which happens to provide normality of the cumulative function. Let us therefore apply the foregoing computational procedure to the case where these conditions of normality and equal step side are met, and observe the effect which size of step has upon the expected distribution of up-and-down trials.

6. CHANGES IN THE EXPECTED DISTRIBUTION OF UP-AND-DOWN TRIALS AS A FUNCTION OF CHANGES IN STIMULUS STEP SIZE, UNDER THE NORMAL MODEL.

Expected distributions of trials are presented in Figure 2 for five different sizes of stimulus step. These distributions are based upon computations by Kappauf and Drucker. The step sizes range from 0.5σ to 3.0σ, where σ is the standard deviation of the cumulative normal.

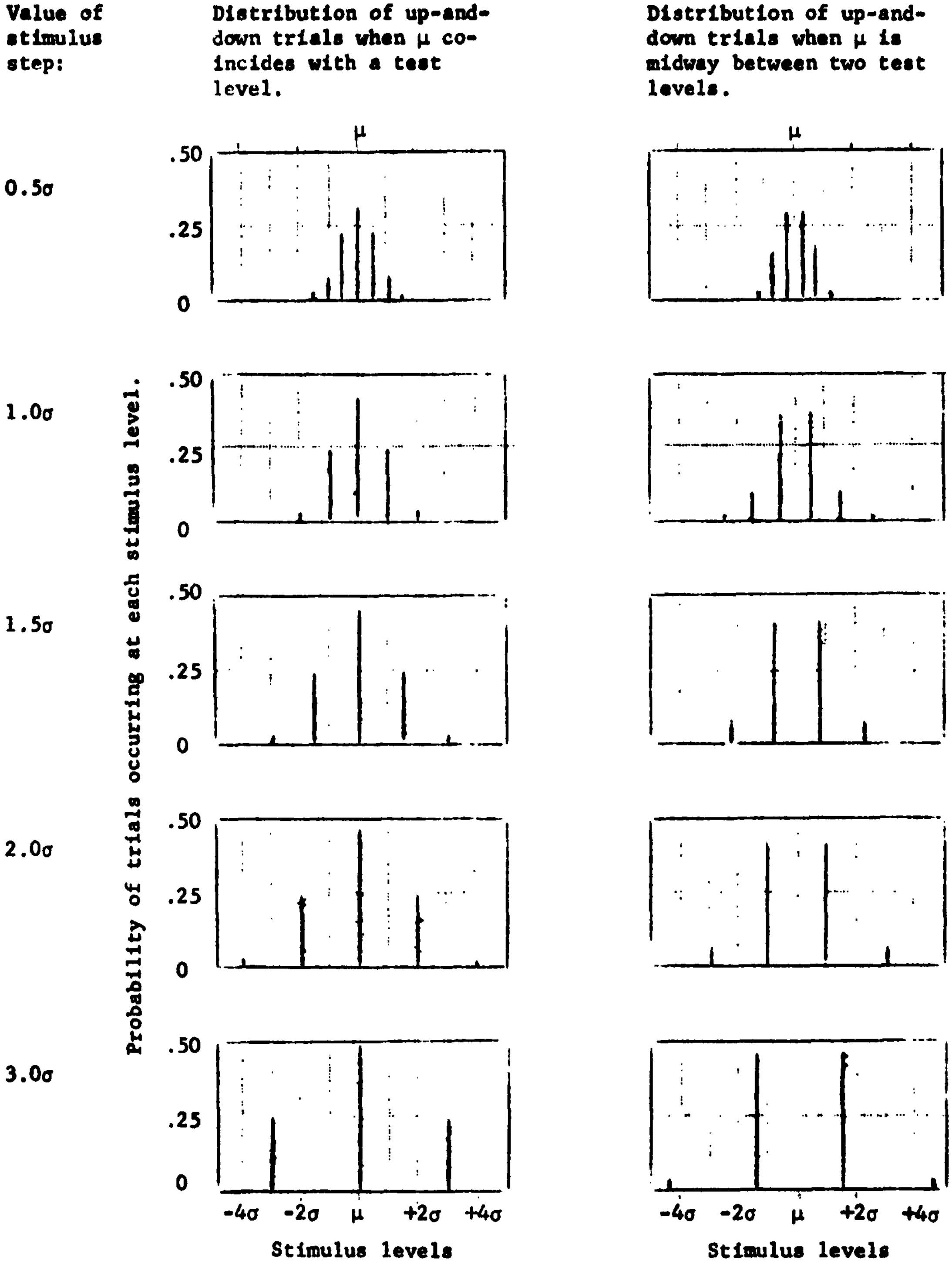

Figure 2: Expected distributions of up-and-down trials under the normal model, for different sizes of stimulus step and for different locations of μ.

Two extreme forms of the expected trial distribution are of interest: one which applies when the mean-median, μ, of the probability of response curve happens to fall exactly at a test level, and the other which applies when the mean-median happens to fall midway between two test levels. (Note: because the mean and median of the cumulative normal are identical, we designate that common value here as the mean-median). Distributions for both of these cases are presented for each of the five sizes of stimulus step. Distributions for other locations of the mean-median would clearly fall between these shown in the figure.

a. Details of computation. In the computations which provided the present distributions, response probabilities at each desired stimulus level were read from the table of the cumulative normal to three decimal places. Values of p in excess of .999 were rounded to 1.000. In the course of plotting the distributions for Figure 2, all probabilities smaller than .005 were dropped. Each distribution is plotted on a common baseline scaled in units of the standard deviation of the cumulative normal.

b. Findings. As we look from the top of the figure to the bottom, we see immediately that when step size is small, i.e. 0.5 σ, the expected trial distribution is such that essentially all trials are conducted within 1.5 σ of the mean-median. As step size grows larger, the expected trial distribution has greater variability until finally when the step is 3.0 σ, trials are conducted with reasonable frequency at stimulus levels 3.0 σ and more from the mean-median.

A second fact which emerges from the figure is that when step size is small, the expected trial distribution makes use of some 6 or 7 stimulus levels, whereas when step size is as large as 3 σ the mean-median is easily bracketed in a few jumps and the expected trial distribution makes use of only some 3 or 4 stimulus levels. (Given a very large stimulus step, of course, only 2 stimulus levels would be used).

A third fact is that the expected trial distribution is not simply the de-cumulated probability of response curve. The expected trial distributions happen to look approximately normal (as may be observed by plotting them on normal probability paper), but they differ considerably among themselves in variance. When the step size is less than 1.0 σ, the variance of the expected trial distribution is less than 1.0 on the scale of σ for the cumulative normal. Similarly, when the step size is greater than 1.0 σ, the variance of the expected trial distribution is greater than 1.0 on the scale of σ.

c. Implications. The foregoing facts have three implications for several sections which follow below. These are:

(1) Re the estimation of the standard deviation of the probability of response curve: The estimation formula will not be based simply on the standard deviation of the obtained trial distribution, because the variability of the latter is not invariant but depends in a critical way upon step size.

(2) Re the effect of skewness of the probability of response curve: The smaller the step size, the greater the probability that the obtained trial distribution for an up-and-down series will cluster around the median and never get out into the tails of the probability of response curve. On the other hand, the larger the step size, the greater the probability that the entire probability function will be scanned, with some trials being conducted at stimulus levels which mark the tails of the function. These facts suggest that whatever the estimate is which is used for the mean-median of the cumulative normal, it will require some special examimation if applied to a situation where the response function is skewed.

(3) Re the difficulty of the subject's differential discriminations: If a small step size is used, the comparison stimulus will almost always be very close to the point of subjective equality, so that the difficulty of the series of trials will be very high. On the other hand, if a large step is chosen, many of the judgments will be easy ones for the subject to make because the comparison stimuli will, in large number, be quite far removed from the PSE.

7. ESTIMATES OF μ AND σ OF THE PROBABILITY OF RESPONSE CURVE, BASED ON THE NORMAL MODEL.

a. The estimates given by the Princeton Research Group. The following estimates are described in a variety of sources: Anon. (1944), Anderson, et al (1946), Dixon and Mood (1948), Dixon and Massey (1951, 1957).

μ and σ are estimated on the basis of the frequency distribution of the responses X, or on the basis of the frequency distribution of the responses Y, whichever response occurred less frequently in the up-and-down series.

Estimate of μ: (i.e. the mean-median)

$$m = \left(\begin{array}{l}\text{mean stimulus level for}\\ \text{the distribution of the}\\ \text{less frequent response}\end{array}\right) \pm .5 \left(\begin{array}{l}\text{value of a}\\ \text{step in}\\ \text{stim. units}\end{array}\right)$$

where:
+ applies if the distribution is for the responses "second stimulus softer".
- applies if the distribution is for the responses "second stimulus louder". (i.e the Jown-moving responses)

Estimate of σ:

$$s = \frac{1.620}{\left(\begin{array}{l}\text{value of step}\\ \text{in stim. units}\end{array}\right)} \left(\begin{array}{l}\text{variance of the}\\ \text{chosen distrib.}\\ \text{in (stim. units)}^2\end{array} + .029 \left[\begin{array}{l}\text{value of}\\ \text{step in}\\ \text{stim. units}\end{array}\right]^2\right)$$

$$= 1.620 \left(\begin{array}{l}\text{value of step}\\ \text{in stim. units}\end{array}\right) \left(\begin{array}{l}\text{variance of the}\\ \text{chosen distrib.}\\ \text{in steps}^2.\end{array} + .029\right)$$

A surprising feature of the estimate for σ is the fact that s depends on the variance of the distribution of trials, not upon the square root of that variance. Note, though, that the second form of the formula above agrees in direction with the facts observed in Figure 2. When the step size is small, the variance of the distribution in step units is large: and when the step size is large, the variance in step units is small. It is reasonable then that these two terms should enter the formula for s as a product. The development of this estimate is described by Dixon and Mood (1948), Anon. (1944). It is an approximate maximum likelihood estimate.

b. <u>The calculation of m (as the PSE) and s (as a measure of the differential threshold) for illustrative data of Figure 1</u>. The preceding formulae are applied in Table 2 to the calculation of m and s for the loudness discrimination data of Figure 1. For these data, computations are based upon the distribution of "s" responses. The symbols A and d are used respectively for the arbitrary origin and for stimulus step. The value of m turns out to be 87.36 db., meaning that the subject's bias was (87.36 db-90.00 db) or - 2.64 db. The standard deviation of the psychometric function is estimated as 1.17 db.

c. <u>The effect on m and s of the position of μ relative to the chosen test levels</u>. The estimate of μ is not seriously affected by the position of μ relative to the test levels, as long as the size of stimulus step is

Table 2

Calculation of m and s from the data of Figure 1.

Judgment Distributions in Figure 1.

Stim. level.	L	S
94	1	
92	1	
90	4	
88	7	3
86	1	7
84		1

1. Use arbitrary origin procedure and work with the distribution having the smaller number of events (The S-distribution here):

Let arbitrary origin = A = 86
Let deviations from A, in step units, be x'
Let the step size be d: here d=2

Stimulus Level	x'	f	fx'	$f(x')^2$
88	+1	3	3	3
86 = A	0	7	0	0
84	-1	1	-1	1
		Σf=N= 11	Σfx'= 2	$\Sigma f(x')^2$= 4

2. Compute m from the formula:

$$m = (\text{mean of chosen distrib.}) \pm \frac{1}{2} d$$

$$m = A + d\left(\frac{\Sigma f x'}{N}\right) \pm \frac{1}{2} d$$

$$= A + d\left(\frac{\Sigma f x'}{N} \pm \frac{1}{2}\right) = 86 + 2\left(\frac{2}{11} + \frac{1}{2}\right) = 87.36$$

3. Compute s from the formula:

$$s = \frac{1.620}{d}\left(\text{Variance of chosen distrib.} + .029\ d^2\right)$$

or

$$s = 1.620\ d\left(\frac{N\Sigma f(x')^2 - (\Sigma f x')^2}{N^2} + .029\right)$$

$$= 1.620\ (2)\left(\frac{11\ (4)\ -\ (2)^2}{(11)^2} + .029\right) = 1.17$$

less than 2.5σ. The estimate of σ, however, is much more sensitive to the position of μ relative to test levels, and there is agreement that up-and-down data do not provide as good an estimate of σ as they do of μ. For small samples, in fact, s may be of little value. How useful it may be in psychophysical research has been explored in some of the studies to be described below.

8. THE ADEQUACY OF m AS AN ESTIMATE OF μ WHEN THE PSCYCHOMETRIC FUNCTION IS NORMAL.

We are concerned here, as we are in relation to any estimate, with the twin problems of bias and reliability.

If an up-and-down series were to consist of a very large number of trials, a very natural estimate to take for μ would be the mean of the stimulus levels used, i.e. the mean of the total trial distribution. It must be noted, however, that the first trial is given at a level which is chosen by the experimenter and has nothing to do with the subject's discrimination. And further, if the experimenter should happen to start the up-and-down series at a stimulus level quite far removed from the subject's PSE, the entire first portion of the series would consist of a succession of like responses which would merely serve to bring the experimenter and the subject into the general vicinity of the PSE. This means that when the up-and-down series consists not of a large number but rather a small number of trials, the simple average of the total trial distribution could be a considerably biased estimate of μ.

One way to think about this matter is in terms of the "expected stimulus level" for each of the early trials of the series. Given a large number of up-and-down series all of which start at a particular stimulus level which is fairly distant from the PSE, the subject's response probabilities at that stimulus level dictate that some of these series will move up for the next trial and some will move down. From this probability information, one can compute the "average" stimulus level at which trial 2 will be conducted. Knowing the stimulus levels which might be used on trial 2 and knowing the response probabilities at each of these levels, one can compute the average or expected stimulus level for trial 3. The computations become more elaborate with each passing trial, but continuing them one may find the expected stimulus level for each of the early trials of the series. One finds, by way of illustration, that if the starting level if 4σ below the PSE and the stimulus step is 1σ, it is not until the tenth trial that the expected stimulus levels is within .01σ of the PSE. Similarly, if the starting level is 4σ below the PSE and the stimulus step is 2σ, the expected stimulus level comes within .01σ of the PSE by the fifth trial of the series. It is the fact that these early expected stimulus levels deviate from μ which would make the average of the total trial distribution a biased estimate of μ.

a. Ways of minimizing bias from early trials. Clearly we wish an estimate of μ which is as unbiased as we can make it. To this end, several procedures have been suggested:

(1) Start the up-and-down series near μ. This is not always possible, particularly in psychological experiments dealing with judgment bias. For such experiments an important advantage of the up-and-down method is that the trial series will hunt for μ, no matter where the series happens to have been started.

(2) Consider the first run of like responses to be a preliminary series of trials which brings the series close to μ. Drop this run from the record and estimate μ on the basis of trials which follow the first "turn-around" or reversal in the up-and-down series.

(3) Consider as many as the first three runs of like responses to be preliminary, and base the estimate of μ on the distribution of trials following, say, the third turn-around.

(4) Consider some fixed number of initial trials as preliminary and drop these from the record. If step size is not too small and if the trial series starts at a level not too far from μ, a reasonable number of initial trials to drop is 5.

(5) Deal only with the trial distribution for the less frequent response, not the total trial distribution. This, of course, is the Princeton procedure cited above. Because the trial distributions for the more frequent and the less frequent responses differ only to the extent that the up-and-down series does not return to the level from which it starts, this procedure has the effect of removing early trials from the computation. Unless the number of trials is large, however, this estimate is still somewhat biased toward the initial testing level (See Dixon and Massey, 1957).

(6) In conjunction with schemes (2) through (5), use double- or triple-size stimulus steps during the trials designated as preliminary. With such steps, the expected stimulus level on successive trials approaches, μ more quickly, turn-arounds occur sooner, and the experimenter obtains early information that the series has bracketted μ.

In the work of our project, we frequently followed procedures (3) and (4), but whether we did nor not for a particular set of data we always used the Princeton estimate, m, for μ (procedure 5 above). The particular programming equipment which we used for the major portion of our work did not permit the adoption of procedure (6), although its use would have been highly desirable. Hopefully, our combined procedures minimized bias.

For a discussion of this bias problem in relation to the efficiency of the estimate for μ, the reader is referred to Brownlee _et_ _al_ (1953).

b. _The standard error of m._ The standard error of m, as we would expect, decreases with an increase in the number of trials in the up-and-down series and increases with an increase in the value of s. It also increases with step size (Anon., 1944; Dixon and Mood, 1948). The Princeton formula is:

Estimate of σ_m: $$s_m = \frac{s\,G}{\sqrt{N_c}}$$

where

N_c is the number of trials in the trial distribution for the less frequent response.

G is the following function of step size.

Step size:	0.5σ	1.0σ	1.5σ	2.0σ	2.5σ
G:	0.94	1.00	1.07	1.15	1.18

Although there was originally some concern that this formula might not provide a satisfactory estimate of the reliability of m, unless the number of trials was of the order of 40 or 50, analyses by Brownlee _et_ _al_ (1953) have shown that this formula is reasonably dependable even when the up-and-down series is very short in length. Our own examination of the variability of values of m for replicated tests using the up-and-down method also confirms the general usefulness of the foregoing estimate of σ_m when the series is of the order of 20 to 30 trials in length, although it appears to overestimate σ_m slightly (see pages 40-42 below).

To the psychologist interested in conducting experiments using relatively short up-and-down series, it may be helpful to have a tabulation of values of

σ_m as a function of the number of trials in the series and the size of the stimulus step. Such a table is presented as Table 3. Note that N_c in the formula for σ_m is the number of trials in the distribution for the less frequent response. $N_{tot.}$ given in Table 3 was taken simply as $2N_c$.

In summary then, we find that regardless of how ppor the experimenter's choice of initial starting level for the up-and-down series, it is possible to obtain values of m which are unbiased estimates of μ. Given that m is unbiased, and given normality and stability of the probability of response function, the standard error of m is reasonably well represented by the Princeton formula.

c. Comparison of the reliability of m with that of the estimate of μ obtained by the method of constant stimuli. It has been shown by Dixon and Mood (1948) as well as by Brownlee et al (1953) that if an estimate of μ is to have a given reliability, this can be achieved with about 30% fewer trials using the up-and down method than by using the method of constant stimuli. This comparison, which is based on the use of the same stimulus steps by the two psychophysical procedures, is of interest in relation to empirical comparisons of the two methods which will be described below.

9. THE MEANING OF m WHEN THE PSYCHOMETRIC FUNCTION IS SKEWED.

In the case where the probability of response curve is normal, the mean and the median of the function coincide. When the function is skewed the mean, μ, moves away from the median in the direction of the longer tail of the distribution. Now it has already been pointed out above that if step size is small, all trials will cluster very close to the median. Oppositely, trials will jump out to test levels represented in the tails of the distribution only if step size is large. The consequence of these relations is that m falls between the median and μ in a skewed distribution. Furthermore, it resembles the median more closely if the step size is small and resembles μ more closely if the step size is large.

Confirming computations of the expected value of m for four distributions of differing degrees of skewness placed this value between μ and the median in all cases. And when the step size was larger, i.e. 2σ, the expected value of m was always closer to the mean of the probability of response curve than when the step size was smaller, i.e. only 1σ. See Figure 3 for

Table 3

VALUES OF σ_m AS A FUNCTJON OF STEP SIZE AND NUMBER OF TRIALS, $N_{tot.}$, IN THE UP-AND-DOWN SERIES, assuming normal model and that initial test level is near μ.

Values of σ_m computed from: $\sigma_m = \dfrac{\sigma G}{\sqrt{\dfrac{N_{tot.}}{2}}}$

Step Size:	Number of trials in the up-and-down series: $N_{tot.}$		
	10	20	30
2.5σ	.53σ or .21 steps	.38σ or .15 steps	.31σ or .12 steps
2.0σ	.52σ .26 steps	.37σ .18 steps	.30σ .15 steps
1.5σ	.48σ .32 steps	.34σ .23 steps	.28σ .18 steps
1.0σ	.45σ .45 steps	.32σ .32 steps	.26σ .26 steps
0.5σ	.42σ .84 steps	.30σ .59 steps	.25σ .49 steps

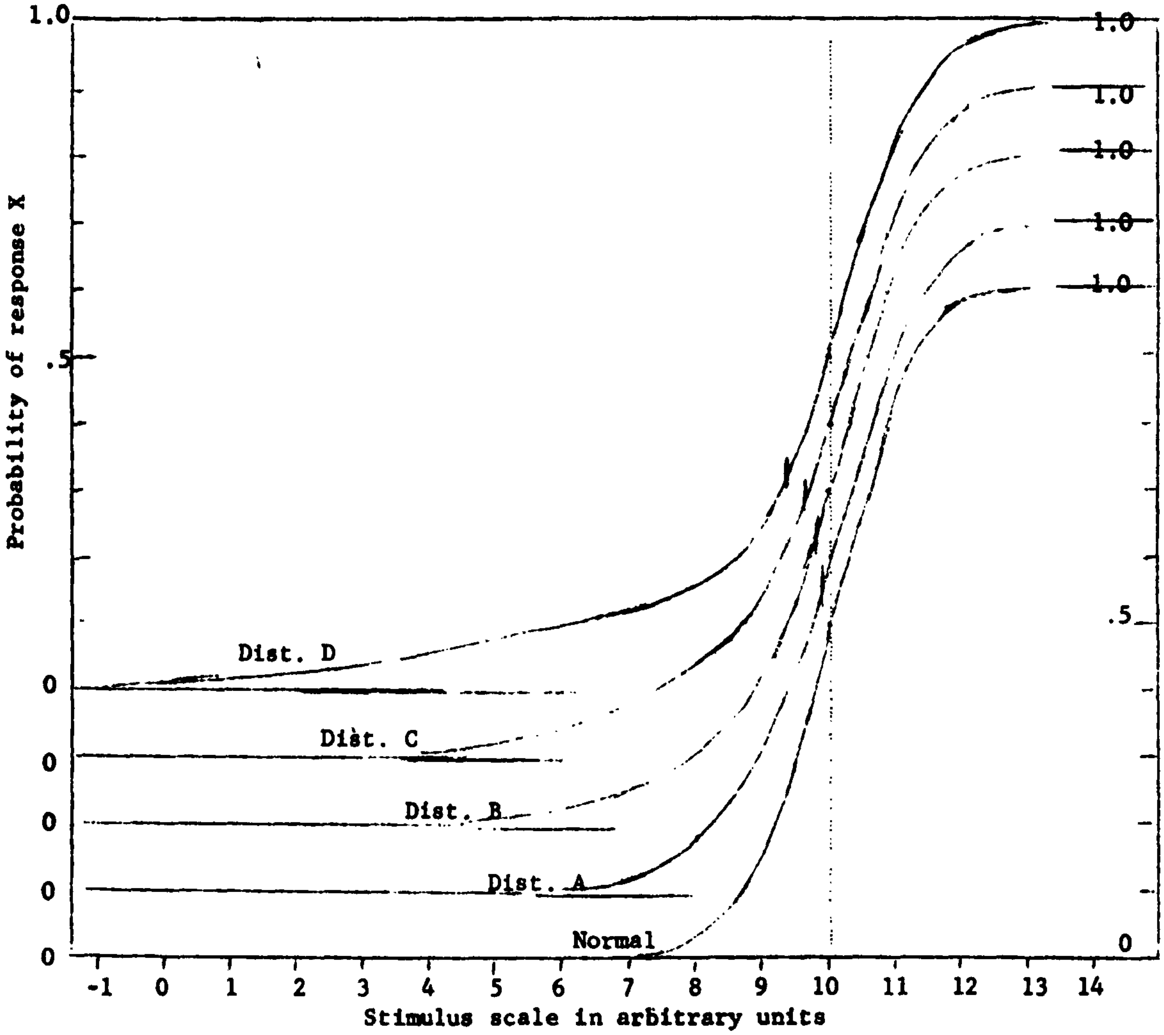

Figure 3: The effect of skewness of the psychometric function upon the meaning of m. The data for the above distributions are given in the table below. All distributions have been plotted above with a common median of 10. The mean for each is shown by the vertical slash.

Distribution:	Normal	Dist. A	Dist. B	Dist. C	Dist. D
Median:	10.0	10.0	10.0	10.0	10.0
m when d=1.	10.0	9.94	9.92	9.87	9.89
m when d=2.	10.0	9.92	9.86	9.79	9.78
m when d=4.	10.0			9.74	9.51
Mean:	10.0	9.90	9.80 ,	9.65	9.40

details. For purposes of these calculations, skewed distributions were devised as composites of two normal distributions. The relative weights given to the component distributions varied, as did the relative size of their variances and the displacement of their means. No specific meaning is to be attached to the scale units in which the distributions are drawn. Stimulus step sizes are given on the same arbitrary scale.

It is clear that the up-and-down method will be applied by psychologists in many measurement situations where in fact the probability function is skewed. In these cases, m will never be far from the "middle" of the distribution, but its expected value will never be either μ or the median of the psychometric function. That this will ever be of critical importance is doubted. Such importance as it may have, however, will depend upon the degree of skewness (i.e. on the difference between μ and the median) and also upon the homogeneity of the sets of data being compared. If all data are for similarly skewed functions, for example, differences between several obtained values of m should still be valid estimates of differences between the corresponding values of μ, or of the medians.

It may be noted that one study of the effect of non-normality on the estimates, m, has been reported in the literature (see Votaw, 1948) but this paper deals with response functions other than those with simple skewness.

10. PREFERRED STEP SIZE; IN σ UNITS.

The importance of step size for the effectiveness of up-and-down testing has already become apparent from preceding sections. We would like a step size, d, which will provide the best possible estimate of μ as well as a good estimate of σ. For psychological experiments, we also want it to be true that the chosen size of step will favor good motivation on the part of our subjects. If compromise on step size is necessary, the following facts are relevant:

a. Reasons for avoiding a step size which is too small.

(1) Too small a step size leads to a waste of many early trials if that step size is used from the very beginning of the series and the experimenter happens to start at a level which is some distance from μ. (See section 8 above).

(2) A step size smaller than 1σ leads to a less reliable estimate of σ than does a step size between 1σ and 2σ. (See Anon., 1944; Dixon and Mood, 1948).

(3) A small step size keeps all trials very near the 50-50 point on the psychometric function. (See section 6 and Figure 2 above). For differential judgments in particular, this means that the subject never has any easy trials. Motivation may fall if he feels that he is guessing on every trial. Note that with a step size of 0.5σ, 70% of all trials are conducted within 0.5σ of μ, and essentially no trials are expected to occur at "easy" levels as far as 2σ from μ. With a step size of 2σ, however, only about 25% of all trials are conducted within 0.5σ of μ, while some 33% of all trials are conducted in the "easy" range, 2σ or more from μ.

b. Reasons for avoiding a step size which is too large.

(1) A step size larger than 2σ makes it necessary to follow a more complicated procedure to estimate σ than that given above. (See Anon., 1944; Dixon and Mood, 1948). As Dixon and Massey (1957) put it, the above estimate of σ is quite accurate as long as the variance of the chosen distribution is larger than 0.3 when computed from data in stimulus steps, but breaks down when the variance becomes less than 0.3 (i.e. $s<.5$).

(2) A step size larger than 2σ leads to a less reliable estimate of σ than does a step size between 1σ and 2σ. (See Anon., 1944; Dixon and Mood, 1948).

(3) For increasingly large step sizes, the value of G in the formula for s_m grows larger and larger, and hence the value of s_m increases. It is helpful to know, however, that over the range of step sizes from 1σ to 2σ, G increases by only a small amount, namely about 15%. (See section 8 above).

(4) For step sizes larger than 2.5σ, G depends upon an unknown value, namely the amount by which μ departs from the testing level nearest to it. This means that for very large steps, s_m is never well determined. In fact, trials using exceedingly large steps may bracket μ repeatedly without providing any information at all about the intervening position of μ.

c. The preferred step size: between 1σ and 2σ. The facts just presented point to the use of a step between 1σ and 2σ. The general recommendation of those who made early evaluations of the method was to use a step as close to 1σ as possible. For one-trial-per-subject experiments (see discussion below) this recommendation stands. For prolonged differential discrimination tasks with the same subject, however, it seems quite essential that we ease the subject's task by giving him a reasonable number of easy judgments. This means using stimulus steps as large as can be tolerated by other considerations. Here we suggest a step size close to 2σ -- say at least 1.5σ, but not exceeding 2σ. The gain in measurement reliability which comes from better subject motivation is not easily quantified, but it probably exceeds the small loss in the reliability of m which comes with the use of steps which are 2σ in size rather than 1σ.

11. METHODS FOR ESTABLISHING A STIMULUS STEP OF THE PREFERRED SIZE.

Knowing how large a step should be in σ units still leaves us with the matter of determining how large the step should be in stimulus units. Two procedures are useful here.

The first and obvious procedure for choosing d is on the basis of an estimate of σ: (a) Run a preliminary series of up-and-down observations using any step thought to be suitable. (b) Compute s for this series. (c) Run a new series with step size between 1.5s and 2s. (d) Compute s for this new series and see if it agrees with the first estimate, etc. From a number of such series the value of σ can be approximated and the stimulus step chosen accordingly.

The other procedure is cruder, may require more data, but has merits of its own as a simple way of monitoring the chosen value of d throughout the experiment. This procedure does not depend upon the calculation of s for each series, but is based instead on the range of test levels used in typical up-and-down series -- a strategy suggested by the distributions in Figure 2, page 15. Suppose that over a number of up-and-down series, no series ever requires the use of more than 3 stimulus levels. The implication from Figure 2 would be that the step size is too large, probably 3σ or more. On the other hand, if as many as 6 stimulus levels are used in series after series, the implication would be that the step size is too small, probably in the range 0.5σ to 1.0σ. These observations led us (Guth and Kappauf) to conduct an empirical sampling study to determine how many stimulus levels would be used with what relative frequency if the stimulus step were 1.5 to 2σ in size. Our finding was that for moderately short up-and-down series, most series would involve 3 or 4 levels, only a few would involve 5 levels, and none would be expected to extend over as many as 6 levels. The specific data obtained for these and other step sizes are shown in Table 4. The information given there may be taken as a guide while one is making the original selection of stimulus step size for an experiment, but perhaps more importantly, can be used to monitor one's choice of stimulus step while the experiment is under way. Changes in the number of test levels used as the experiment proceeds may reflect practice effects, changing test conditions, etc. and point to the desirability of revising the value of d for testing under these new conditions. The general rule which emerges from the table is that our stimulus step is in the range of 1.5 to 2σ if, starting at an initial test level near μ, we find the frequent use of 4 stimulus levels, the infrequent use of 5, and no use of 6.

(One side comment is in order here. The mere fact that testing over a series of stimulus levels results in a see-sawing up-and-down series is in itself no proof that behavior is changing with stimulus level or that the probability of response function over these levels ranges from .00 to 1.00. Suppose that the probability of response X and the probability of response Y are each .50 at every stimulus level. The up-and-down series then becomes an illustration of the mathematician's random-walk problem. Empirical sampling under these conditions indicates that 15-trial series will range on the average over 6.0 test levels, 20-trial series over 7.3 test levels, and 30-trial series over 8.5 test levels. Such series might appear to be bracketting a 50-50 point, but of course they are not. The use of a large number of test levels may thus mean, in certain untried test situations at least, that there is nothing to measure, no 50-50 "point" to locate.)

12. THE REQUIREMENT OF INDEPENDENCE OF SUCCESSIVE TRIALS, AND WAYS OF MEETING IT, IN WHOLE OR IN PART.

Thus far in our discussion, little attention has been given to the fact that the model for the up-and-down method presumes that the probability of response curve is unchanging from trial to trial. The probability of each of the alternative responses at a given stimulus level is presumed to be the same regardless of the nature of the previous trial and the response which was made on that trial. This is to say that successive trials are assumed to be independent. All the computations of expected trial distributions were based upon this premise.

Table 4

Results of empirical sampling study: probability of using a given number of test levels in an up-and-down series as a function of step size and number of trials ($N_{tot.}$), assuming normal model and that initial test level is near μ.

Step Size	Number of test levels used	μ exactly at a test level, and $N_{tot.}$ (the number of trials in the series) is				μ midway between test levels, and $N_{tot.}$ (the number of trials in the series) is			
		15	20	25	30	15	20	25	30
3.0σ	3 or less	1.00	1.00	1.00	1.00	.88	.87	.79	.75
	4					.12	.13	.21	.25
	5								
	6 or more								
		(Mostly 3's, no 5's)							
2.0σ	3 or less	.85	.80	.75	.75	.62	.43	.42	.20
	4	.13	.13	.17	.15	.38	.57	.58	.80
	5	.02	.07	.08	.10				
	6 or more								
		(Frequent 4's, some 5's)							
1.5σ	3 or less	.48	.37	.38	.30	.42	.20	.25	.15
	4	.50	.60	.50	.55	.58	.80	.75	.85
	5	.02	.03	.12	.15				
	6 or more								
		(Very frequent 4's, some 5's)							
1.0σ	3 or less	.25	.20	.17	.10	.20	.10	.00	.00
	4	.68	.53	.50	.45	.60	.57	.58	.50
	5	.08	.27	.33	.45	.20	.30	.38	.45
	6 or more					.00	.03	.04	.05
		(Frequent 5's, some 6's)							
0.5σ	3 or less	.05	.03	.00	.00	.05	.00	.00	.00
	4	.55	.50	.33	.25	.55	.57	.38	.20
	5	.30	.33	.38	.35	.25	.23	.38	.45
	6 or more	.10	.13	.29	.40	.15	.20	.25	.35
		(Frequent 6's)							

Note: The probabilities cited in this table were determined from 10 up-and-down series of 600 "trials" each, where the outcome of each trial was determined by consulting a table of random normal deviates. The 10 series involved the 5 step sizes and the 2 locations of μ. Each series of 600 trials was then subdivided into sub-series of 15, 20, 25 and 30 trials on which counts of the number of levels used were made. Thus, the probabilities for series of length 15 were based on 40 sub-series, those for series of length 20 on 30, of length 25 on 24, and of length 30 on 20.

a. The up-and-down series based on one trial per subject. One way of assuring the independence of successive trials is to conduct each trial on a different individual. In this case, the probability of response curve applies to the population of individuals, and the variability of this curve reflects both intra- and inter-individual variability. The procedure is nevertheless effective for some studies, and we have used it in work on taste preferences in animals as well as in the experiment on the time error with human subjects. These applications will be discussed below.

b. Concurrent up-and-down series with same subject. Most often, however, the object of a psychophysical experiment is to quantify the discrimination of individual subjects. To this end the same subject must be tested over many trials. Under these circumstances there are at least three forms of trial dependencies which could occur.

The first of these is dependency associated with the direct effect of the immediately preceding trial(s) upon each current judgment. This, of course, is a matter of experimental interest in its own right, and is THE matter of experimental interest for the context studies reported later.

The second form of possible trial-to-trial dependency is that associated with limen drift, limen fluctuation, criterion fluctuation, etc., during the course of the experiment (see Day, 1951; Verplanck et al, 1952; Kappauf and Payne, 1954, 1955). If such drifts or fluctuations occur, we have a situation where μ is really changing and not remaining fixed. Successive trials are not completely independent because they sample the subject's behavior at moments when μ has varied little if at all, whereas more widely separated trials occur at times when μ may be different.

The third form of trial-to-trial dependency is unique to sequential testing methods and has long been a subject of discussion with regard to the method of limits (see e.g. Titchener, 1905; Urban, 1908). It concerns the fact that if the subject is aware of the sequential order of trials, his expectation or set will change as the series of trials proceeds. For the up-and-down method, in particular, a subject might well "see through" or "discover" the experimenter's up-and-down program of trials. Once he does so, he might reduce his judgments on successive trials to a simple alternation between response X and response Y. Clearly successive trials would no longer be independent.

There are several ways in which we should be able to conceal the up-and-down trial sequence from our subject, and hence deal with this third source of trial dependency. One way would be to intersperse one or more "dummy" trials between successive "regular" trials of the up-and-down series. (This corresponds to Fernberger's procedure (1913) for obscuring the serial order of trials in the method of limits). Another way would be to arrange the interspersed trials as trials on other up-and-down series. The latter plan means that we conduct the epxeriment with a number of concurrent up-and-down series. Successive trials in the experiment are then randomly chosen from the different series. Experience indicates that this completely conceals the basic testing routine from the subject. In our experiments, the number of concurrent series used in tests with a single subject has ranged from 3 to as many as 12.

Suppose that several concurrent series are run all using the same standard. Here one would anticipate that the effects of preceding trials would be random, and hence equivalent within random error, for each of the several series. Limen drift or criterion drift, which ought to affect all series about equally, should tend to increase s for each series but stabilize the values of m for the several series. Thus the values of m may differ somewhat less from one another than would be expected on the basis of some composite estimate, say $\bar{s}_m$, of σ_m. This problem will be examined in more detail in section 14 below with regard to some loudness discrimination data.

The use of several concurrent up-and-down series during continued observation by a given subject is thus thought to eliminate the potentially most troublesome source of trial-to-trial dependency but leaves two others. Both of the latter will be the subject of study below.

13. UP-AND-DOWN TESTING BASED ON ONE TRIAL PER SUBJECT: RESULTS OF FOUR EXPERIMENTS AND OBSERVATIONS ON THE VARIABILITY OF m OVER REPLICATIONS.

The one-trial-per-subject procedure is of interest because it guarantees independence of successive trials in the up-and-down series, and also because it rules out bias which, under a massed-trials-with-the-same-subject procedure, may arise from the subject's continued experience in the test situation. Such bias is found in context effects from comparison stimuli, in effects associated with changes in motivation level with continued testing, etc.

We first used the one-trial-per-subject procedure in an early project experiment where it was our interest to measure the time error under conditions which would be as free as possible from context effects. Subsequently the procedure was used in a series of non-project experiments designed to map families of taste preference functions (isohedons) for the white rat.

Here our concern was to limit the quantity of incentive fluids ingested by any one animal in order that the obtained preference functions would be associated with taste preferences as such, uncontaminated by the effects of post-ingestional factors.

Clearly, the one-trial-per-subject procedure can be efficient if and only if the instruction time or "readying time" per subject is not long. This condition was met in the applications to be discussed here.

a. Experiment T-1: The effect of inter-stimulus interval on the magnitude of the time error, measured under conditions of minimum context.

(1) Introduction: The "time error" (or the "time order error") is the name given to that bias in differential judgment which arises as a function of the fact that the two stimuli to be compared are separated in time. Most often the time error proves to be negative, that is, the second stimulus need not be as intense as the first in order to appear equal in intensity to the first. Under some conditions of stimulus context or of inter-stimulus interval, however, the error proves to be positive, that is, the second stimulus needs to be more intense than the first if the two are to appear equal in intensity. Problems which have attracted interest with regard to the time error in differential judgment include the following: (1) How does the time error vary in magnitude as a function of inter-stimulus interval? (See Kohler, 1923; Needham, 1934, 1935; Koester, 1944-45). (2) How general is the time error? For what judgment situations is the time error present, for which ones absent? (e.g. Postman, 1946; Stevens, 1957). (3) How does the time error for judgments with a given standard vary as a function of stimulus context, i.e. as a function of having trials with other standards scheduled in the same session? (e.g. Woodrow, 1933; Needham, 1935; Koester and Schoenfeld, 1946). (4) How may one best explain the time error and the manner of its variation with situation variables? (Kohler, 1923; Pratt, 1933; Woodrow, 1933; Michaels and Helson, 1954).

In general the time error has been a very variable phenomenon, not only variable from individual to individual (Needham, 1934) but especially variable as a function of experience in the test situation (Kohler, 1923; Needham, 1934; Stott, 1935). Continued experience, however, has always meant continued exposure to the same set of comparison stimuli as called for by the method of constant stimuli. It would appear then that context factors may have been important in determining the time errors displayed by experienced subjects. One time error function which is reported to be very sensitive to experience effects is the so-called "p-function," i.e. the function relating the time error to inter-stimulus interval. (Kohler, 1923; Needham, 1934). Because this function has been central in some theoretical discussions of the time error, it appeared of interest to determine the effect of inter-stimulus interval of the time error for completely naive subjects - subjects who are tested for only one trial each.

In using the up-and-down method in this study, we gained two advantages. (a) We were able to specify the magnitude of the time error in stimulus units. Previous work had had to rely on less direct measures as the relative proportion of preponderance of "greater-than" and "less-than" judgments (i.e. the D-% measure). (b) We were able to evaluate our results using significance tests based upon computed values of s and s_m.

(2) Purpose: to determine, for several discrimination tasks, the magnitude of the time error as a function of inter-stimulus interval under conditions of minimum context, i.e. where each subject makes only one judgment per discrimination task.

(3) Discrimination tasks: Each subject made three differential judgments in this order: a judgment of the difference between two weights, one of difference between two auditory durations, and one of the difference between two pressures. The stimuli employed were the following:

Lifted weights: The subject lifted 50 cc. Erlenmyer flasks filled with cotton and lead shot. The standard was 100 grams. The comparison stimuli were in the logarithmic series 71, 79, 89, 100, 112, 126, etc. grams.

Duration discrimination: The stimulus employed was a buzzer controlled in its duration by a Stoelting timer. The standard was of 3 seconds duration. The comparison stimuli were in the logarithmic series 2.13, 2.25, 2.38, 2.52, 2.67, 2.83, 3.00, 3.18, 3.37, 3.57, 3.78, 4.00, etc. seconds.

Pressure discrimination: The subject depressed a key with his forefinger. The key was on one end of a weighted lever and the subject could experience the pressure by depressing the key a distance of one-eighth inch. The pressure was 50 grams for the standard, and was 28.1, 31.6, 35.4, 39.7, 44.5, 50.0, 56.2, 62.8, 70.7, etc. for the comparison stimulus conditions.

Inter-stimulus intervals: These were 2.5, 5.0, 10.0 and 20.0 seconds.

(4) Procedure: Each subject came to the laboratory individually. He sat at a table where he could perform all three tasks. The subject was shown the 100 gram standard, instructed in lifting it for 3 seconds, and allowed to lift it once for practice and for familiarizing himself with its weight. He was told that our purpose was to see how well he could judge

the difference in weight between two flasks when the second was not available until __ seconds after the first had been put down. The trial was run with the subject lifting on the word "LIFT" and lowering the weights on the word "DOWN." He judged the second weight as "Heavier" or "Lighter" than the first. He was then introduced to the duration task and allowed to listen once to the standard so that he would know what general length of sound to expect. He was told what the delay would be between the two buzzes when the trial was conducted, and then the trial was run with judgments of "Longer" or "Shorter" being the only ones admissable. For the pressure task the routine was similar, the subject being allowed to depress the key to feel the standard once before the comparison trial began. In the trial itself, he depressed the key on "Down" and released it slowly on "Up." Again he was allowed only judgments of "Heaver" or "Lighter." Three different delays were used for the three tasks for each subject. The order of delays was balanced across subjects so that at the end of the experiment 45 subjects had made each judgment with each delay.

(5) Subjects: The subjects were 180 undergraduates, men and women, recruited from classes in elementary and experimental psychology.

(6) Results: The data are summarized in Figure 4 and Table 5.

For the lifted weights discrimination, the values of s for the four delay conditions were not significantly different when evaluated by Hartley's F_{max} test. The weighted average value of s, $\bar{s}$, was 1.16 steps, implying that the step size, d, was not too badly chosen for this task, being about 0.9 $\bar{s}$. (It had been our objective to obtain steps approximately equal to 1σ for all three discrimination tasks)., s_m computed from $\bar{s}$ proved to be .25 steps. The range of the four values of m, i. e. the time error, for the different delay conditions was tested against s_m (.05 level q-test). This test led to rejection of the hypothesis that the inter-stimulus interval was without effect upon the time error. Based on the amount of data here collected, the time errors for the 2.5 and for the 5 second inter-stimulus intervals were not significantly different from zero (.05 level tests), while those for 10 and 20 seconds were. The overall pattern of the results plotted in the figure, however, are consistent with classical results: the time error for short inter-stimulus intervals, say below 3 seconds, appears to be positive for naive subjects.

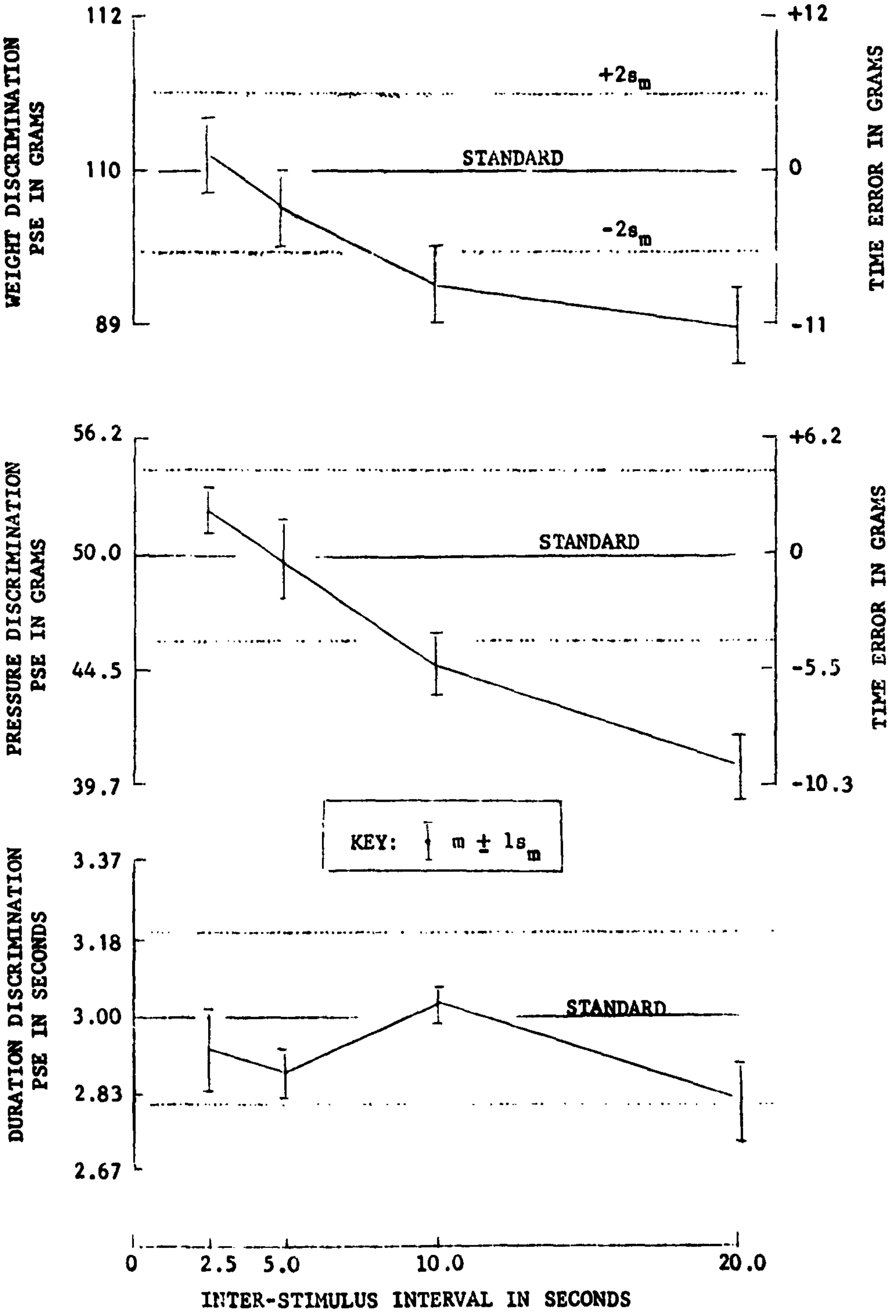

Figure 4: PSF and time error as a function of interstimulus interval for three discrimination tasks.

Table 5

DATA FOR EXPERIMENT T-1: TIME ERROR AS A FUNCTION OF INTER-STIMULUS INTERVAL.

WEIGHT DISCRIMINATION (100 gm standard)	2.5 secs	5.0 secs	10 secs	20 secs
PSE in steps *re* standard	+0.09	-0.23	-0.73	-1.00
PSE in grams	101.0	97.4	91.9	89.0
Time Error in grams	+1.0	-2.6	-8.1	-11.0
s in steps	1.32	0.81	1.07	1.34
s_m in steps (based on $\bar{s}$)	0.25			
PRESSURE DISCRIMINATION (50 gm standard)				
PSE in steps *re* standard	+0.40	-0.07	-0.93	-1.74
PSE in grams	54.4	49.6	44.8	40.9
Time Error in grams	+2.4	-0.4	-5.2	-9.1
s in steps	.80	2.14	1.22	1.42
s_m in steps (conservative)		0.38		
DURATION DISCRIMINATION (3 secs. standard)				
PSE in steps *re* standard	-0.41	-0.69	+0.14	-1.05
PSE in secs.	2.93	2.88	3.02	2.82
Time Error in secs.	-.07	-.12	+.02	-0.18
s in steps	3.27	1.68	1.31	2.95
s_m in steps (conservative)	0.56			

For the pressure discrimination, the function obtained agreed well with the pattern of the function obtained for the weight discrimination. This agreement should not be too surprising in view of the kinesthetic similarity of the two tasks. In the case of pressure, the values of s for the up-and-down series for the four different time intervals were significantly different (.05 level, F_{max} test). The average step size was smaller than desired, being about 0.7 s. The most conservative estimate of the value of σ_m, based on the largest obtained value of s, was .38 steps. In terms of this value, the four obtained time errors differed significantly (.05 level, q-test) and the errors at 10 and at 20 seconds were significantly different from zero. Note that the time errors were relatively larger for the pressure discrimination than for the weight discrimination. Nevertheless, the influence of inter-stimulus interval on time errors for the two discriminations was very similar.

For the duration discrimination, the results were far less neat and clear. We believe this to have been a function of the fact that the step size, which had been chosen on the basis of pilot work with experienced subjects, proved to be for naive subjects considerably smaller than desired, namely about 0.4 s. For such a step size, s has fairly low reliability. Analysis of the data indicates that the values of s obtained for the four different delay conditions were significantly different, while the values of m (the time errors) for the four conditions were not significantly different (conservatively tested using largest obtained s as basis for error estimate). The over-all time error, averaged over the four inter-stimulus conditions was about 0.5 steps, and the data do not justify concluding that this was significantly different from zero. The obtained error was negative, however, and this does agree in direction with results obtained by Stott (1935), and by Woodrow and Stott (1936) for duration judgments when the standard is of 3 seconds duration.

(7) Discussion. Two methodological comments are in order. First: it is of interest to note that since the testing of any one subject in the present experiment did not require an extended series of trials, it was not inconvenient to use inter-stimulus intervals as long as 20 seconds. Most studies in the literature had stopped short of this long a delay interval. Second: the duration portion of the study points to the importance of having the size of stimulus step within the desired range, as discussed in Section 10 above. Were the duration tests being conducted

today, we would benefit by the sampling data presented in Table 4, page 30 above. The up-and-down series for the first 15 subjects had already covered 5 stimulus levels in the case of the 2.5 second delay condition, 5 for the 5 second condition, 5 for the 10 second, and 6 for the 20 second condition. With Table 4, we would have been warned that we were using a step size of the order of .5σ and would have revised our stimulus series.

Three content results with regard to the time error are also worthy of comment. First: the time error for completely naive subjects, making one judgment only and hence not subject to context effects arising within the experiment itself, changes for both weight discrimination and pressure discrimination as a function of inter-stimulus interval. The nature of this function is much as Kohler described it on the basis of his early experiments. Second: the magnitude of the time error, which was readily quantified using the up-and-down method, became as large as 11 grams when the 100 gram standard weight had been lifted, and as large as 9 grams when the 50 gram pressure had been experienced. Both of these extreme errors were observed with the 20 second inter-stimulus interval, and it remains a possibility that the error would grow still more of the inter-stimulus interval were extended. Third: there was no tendency for variability of judgment (σ) to increase with inter-stimulus interval, as might have been expected from a variety of points of view.

b. Three experiment to map taste isohedons in the rat.

(1) Background. Some three years ago, P. T. Young became interested in the problem of identifying incentive solutions which were equally acceptable to the rat. Such solutions he called isohedonic, following the lead of Guilford (1954) who had used the term with respect to auditory stimuli which the human subject found equally pleasant. On the basis of our project work with the up-and-down method, we proposed that an effective way to locate a solution mixture which was isohedonic with a given standard solution (or mixture) would be to use a group of 25 to 30 animals and follow the up-and-down, one-trial-per-subject procedure.

(2) Method. Each animal in the group was given a brief, usually 3-minute, preference test in which two solutions were available--a simple sucrose solution which was the standard, and a comparison mixture containing, say, quinine and sucrose. If the first animal licked more of the standard than of the comparison, the next animal was tested with a comparison containing "one step more" sucrose. But if the first animal licked

more of the comparison, the second animal was tested with a less palatable comparison, i.e. one containing "one step less" sucrose. Similar moves were made after each other animal was tested. The animals in the experimental group were tested in a random order, their responses providing a group up-and-down series which hunted about that sucrose level which made the comparison isohedonic with the standard. On each new test day, a different measure was taken, with a different concentration of sucrose in the standard and/or with a different concentration of quinine in the comparison mixture. Each day the animals were tested in a new random order, each one again being used for but one trial on the up-and-down series.

(3) Results. This procedure has been followed in a series of three experiments. Christensen (1962) has worked with sucrose vs. sucrose-salt mixtures. Kappauf, Burright and DeMarco (in press, 1963) have worked with sucrose vs. sucrose-quinine mixtures, and Young and Schulte (in preparation) have worked with sucrose vs. sucrose-acid mixtures. In each experiment it has been possible to locate a series of mixtures all isohedonic with the same standard and thus map complete isohedonic contours or isohedons. The one-brief-trial-per-subject procedure makes it quite certain that these isohedons are descriptive of the animals' taste preferences and not a function of post-ingestional factors which, under other test conditions, might have influenced the animals' choice behavior.

c. An empirical check on s_m as a predictor of variability in m in these experiments. In each of the foregoing animal experiments there were some measurements which were repeated or replicated on a later test day. The size of the differences in m from the first occasion to the second, or more specifically the root-mean-square of these differences may be estimated from $s_{(m_1-m_2)}$, as computed from typical values of s_m and s. Table 6 provides a comparison of this measure of expected variability in m for each of the experiments, with the variabilities actually obtained.

Included in the table, along with the animal data, are similar computations based upon records for the weight discrimination part of the time error study. Here there were no formal replications, but we report the outcome of dividing the series of 45 observations for each delay condition into two portions, the first 20 trials and the last 25 trials. Values of m were computed for each of these "halves," and then the first-half-second-

Table 6.

EVALUATION OF s_m AS A PREDICTOR OF VARIABILITY IN THE VALUE OF m IN REPLICATED ONE-TRIAL-PER-SUBJECT EXPERIMENTS.

EXPERIMENT	COMPUTATION OF EXPECTED VARIABILITY IN m ACROSS REPLICATIONS					REPLICATION DATA	
	Basis for estimate of s	s, in steps	Typical N_c in series where re-tests were made*	s_m, in steps	$s_{(m_1-m_2)}$ in steps	Root mean square difference, in m between replications, in steps	No. of paired (replicated) series
Experiment T-1, Weight Discrim.	$\bar{s}$ as average s for 8 half-series.	$\bar{s}$ = 1.14	11*	.35	.49	.37	4
Christensen: sucrose-salt study with rats.	via range of stimulus levels and Table 4, for case of $N_{tot.}$ = 30.	s 1.0	14	.37	.38	.35	17
Kappauf, Burright and DeMarco: sucrose-quinine study with rats	$\bar{s}$ as average of values of s for the 8 up-and-down series involved in the replications.	$\bar{s}$ = .90	9	.30	.42	.36	4
	$\bar{s}$ as average for all series in the study.	$\bar{s}$ = .83	9	.28	.40		
Young and Schulte: sucrose-acid study with rats.	$\bar{s}$ as average of values of s for the 8 up-and-down series involved in the replications.	$\bar{s}$ = .88	15	.23	.32	.31	4
	$\bar{s}$ as average for all series in the study.	$\bar{s}$ = 1.00	15	.26	.36		

*N_c, to be distinguished from $N_{tot.}$, is the number of observations of the less frequent response, i.e. the N used in the formulae for m, s and s_m.

half difference in the value of m was determined for each of the four time delays. Values of s for the 8 half-series were homogeneous and were averaged to provide $\bar{s}$.

It should be noted that the values of $\bar{s}$ given in the table are simple averages, and hence are smaller than the values which would have been obtained if values of s^2 had been averaged and the root taken. In the case of each of the last two studies cited in the table, two estimates of s are given: one based upon the variability of all series run in the experiment, and the other based only upon the variability of those up-and-down series involved in the replications.

Examination of the second and third columns from the right in Table 6 indicates that the obtained replication differences were closely approximated by the error formula. But it will be seen that for each of these four studies, the value of $s_{(m_1-m_2)}$ was larger than the observed root-mean-square change in m. We infer that the variability of m, as estimated from s, is over-estimated. This result is consistent with a conclusion drawn by Brownlee _et al_ (1953) from their analysis of the up-and-down method with small samples. The variance estimation formula, s_m, is based on asymtotic theory, and the validity of testing procedures using s_m for finite samples went uninvestigated during early work with the method (Anon, 1944). What Brownlee _et al_ subsequently observed was that s_m may provide a conservative estimate of the accuracy of m for small samples when up-and-down series start close to μ, say within 2 testing levels of μ. Such "close starts" did characterize the up-and-down series in the present experiments.

We thus have good reason to believe that for many, perhaps most, applications of the up-and-down method in psychological research s_m will be a conservative estimate of the accuracy of m when the one-trial-per-subject procedure is employed. This conservative feature of s_m and the adequacy of tests based on s_m are clearly deserving of continued study. (Please note in this connection that the tests cited on pages 35 and 38 above were conducted taking s_m at its face value).

d. Summary comments. For experiments where group average performance is to be evaluated, and where readying time per subject is suitably brief, up-and-down testing following the one-trial-per-subject procedure may frequently prove useful. One-trial-per-subject experiments have seen limited applications in the past but they have been used (e.g. Stevens, 1956) and should be used when it is of importance to eliminate the effect of the subject's having had other recent experience in the test situation.

The estimate, s, obtained from the up-and-down series appears to provide a satisfactory, though somewhat conservative basis for estimating the reliability of m for simple replications with the one-trial-per-subject design.

14. UP-AND-DOWN TESTING WITH CONCURRENT SERIES RUN UNDER THE SAME TEST CONDITIONS AND WITH THE SAME SUBJECT: THE VARIABILITY OF m FOR CONCURRENT SERIES.

When several concurrent up-and-down series are run using the same standard, the same test conditions and always the same subject, the values of m obtained for these separate series each estimate the same parameter, μ, the subject's PSE. Variability among these estimates will be influenced by a number of factors: first, by s, G, and N_c, as we know from the formula for s_m; second, by differences between the initial testing levels used for the different series, in the event that some or all initial levels are far enough from μ to bias m; and third, by whatever serial dependencies may exist among the judgments or responses of the session-long program of trials.

a. Effect of differences in initial testing level on the values of m for concurrent series. We have already discussed the general problem of bias in m associated with the starting level of an up-and-down series, and have considered ways of minimizing this bias (see pages 20-22 above). From that discussion we recognize that if our several concurrent series (or the portions used in computing values of m) all start near μ, the bias will be small for each, and so the influence which differences in starting level will have upon the values of m for the different series must also be small. We anticipate, therefore, that concurrent series may be managed so that the variability of m will not be enhanced by differences in initial testing level.

In connection with the analysis of the data of two loudness discrimination experiments (L-15 and L-16 which will be described below), McDiarmid (1962a) had occasion to pull out a sum of squares associated with initial testing level. In these experiments, there were many cases where two relatively short concurrent series, with fewer than 25 trials per series, were run with the same standard. These series started 2 stimulus levels apart, at levels which were the same for all subjects in a particular test group. PSE's were different for different subjects, of course, but the average distance of the PSE from the more distant starting level was only some 2 or 2-1/2 stimulus steps. These conditions should introduce little bias in the values of m, and indeed McDiarmid could find in his analysis no basis for rejecting the hypothesis that initial testing level had been without effect upon m.

b. <u>Dependecies between successive observations</u>. Some exceptionally long concurrent series were run in a portion of one of the foregoing loudness experiments (L-16, Part I for Group R-1). The records of these series permitted a cross-correlation study of dependencies between successive observations in the test session (See McDiarmid, 1962a).

This part of the experiment involved four concurrent series. Trials were run taking the four series in simple rotation: one trial from series #1 for the less intense standard, one from series #1 for the more intense standard, one from series #2 for the less intense standard, one from series #2 for the more intense standard, etc. The number of trials per series was 75. Context effects had stabilized by the end of 25 trials per series, so the last 50 trials per series were appropriate for cross-correlational analysis. Because of the influence of range on r, it seemed desirable that this analysis be carried out with the more variable of the available series, i.e. with series where the chosen step size had been relatively smaller. This directed our attention to the two series for the less intense standard.

Of the final 200 trials in the test session, trials 101, 105, 109, . . . 297 belonged to one series with the less intense standard, and trials 103, 107, 111, . . . 299 belonged to the other. For purposes of computing a cross-correlation between the stimulus levels used in these two series, the paired trials might be taken as 101 and 103, 105 and 107, etc. or they might be taken as 103 and 105, 107 and 109, etc. It is clear that there is no reason to prefer one of these pairings to the other, and that each pairing produces a biased value of r depending upon the phase relationships between the two up-and-down series. (This bias reaches its limit when the step size is very large and μ is between two levels. Then each series oscillates back and forth between two test levels, one on either side of μ. Under these circumstances, one pairing of the trials produces an r of +1.00 and the other an r of -1.00). We therefore took the average of the correlations found for the two different cross-pairings of trials as our measure of the cross-correlation of the two concurrent up-and-down series. The data for the 8 subjects who served in the group under consideration given in Table 7.

As expected from its nature, r varied with the size of s. The up-and-down series for the most variable subject drifted considerably during the trials under study, and the correlation measure for him shows that indeed the two series generally moved up and down together. Some of the other correlations are small, but it will be noted that seven of the eight average cross-correlations are positive.

The tendency found here for concurrent up-and-down series on the same standard to drift up and down together adds to accumulating data in the literature which point to trial-to-trial serial dependencies. It is noteworthy here that the present dependencies were observed at a lag of 2 trials, one trial with the more intense standard having intervened between the members of every pair of trials which entered into the cross-correlations. Presumably the cross-correlations for two concurrent series taken in simple alternation on successive trials would be greater than those cited in Table 7.

c. The variability of values of m for concurrent series run under the same test conditions. From the foregoing correlations we may presume that fluctuations of the subject's criterion of equality or changes in his response habits occur during the test session. Such changes must have the effect, as suggested on page 32 above, of raising the variability observed within the single up-and-down series, while at the same time stabilizing to a certain extent values of m for the different series. This then is a factor which will cause the values of m for concurrent series run under the same conditions to vary less than we would expect on the basis of obtained values of s and s_m.

Table 7

CROSS-CORRELATIONS BETWEEN CONCURRENT UP-AND-DOWN SERIES FOR 8 SUBJECTS IN A LOUDNESS DISCRIMINATION EXPERIMENT

(Data for Group R-1, Experiment L-16, Part I, 53 db. standard)

Subject	Average value of s for 2 concurrent series: $\bar{s}$	Step size in units of s	Cross-correlations		
			for one phasing of trials	for other phasing of trials	Average
1	5.10 steps	.20s	+.53	+.54	+.53
2	0.97 steps	1.03s	+.13	-.10	+.01
3	1.86 steps	.54s	+.07	+.05	+.06
4	1.66 steps	.60s	+.09	+.21	+.15
5	1.19 steps	.84s	.00	-.06	-.03
6	1.94 steps	.52s	+.26	+.24	+.25
7	1.44 steps	.69s	+.08	+.26	+.17
8	1.12 steps	.89s	+.28	+.16	+.26
Average	1.91				+.17

We have already seen, of course, that s_m is a conservative estimate of the reliability of m in the one-trial-per-subject situation, so the occurrence of positive cross-correlation between concurrent series means that s_m should be an even more conservative estimate of the reliability of m for concurrent-series.

McDiarmid (1962a), continuing the analysis of the data which provided the correlation measures in Table 7, made further calculations to compare the average value of s for the 8 subjects in the group with an estimate of σ based on the observed variation in the value of m from one concurrent series to the other. In Table 8 we present a summary of his calculations, recast in a form which compares expected and observed concurrent-series differences in m. This table is thus a direct parallel of Table 6, page 41. The comparison on interest here is between the last two columns at the right. Again, as in Table 6, s_m leads to an overestimate of the root-mean-square difference in the obtained values of m. For the less intense standard, as compared with the more intense standard, the average value of s was larger (step size smaller), and it was for this reason that the correlations in Table 7 were computed for the less intense standard. Cross-correlations for the more intense standard were not calculated, but they must have been less than those cited in Table 7. Were they close to zero, the observed and expected root-mean-square concurrent-series difference in m should have differed by an amount similar to the discrepancies found in Table 6. And this is the case for that standard. For the less intense standard, however, the discrepancy is greater than any reported in Table 6, a result associated with the observed cross-correlations and interdependencies.

It would appear from these records that with step sizes in the range thought to be desirable, namely 1.5 to 2.0σ, the effect of cross-correlation on the variability of m will be neglible: that for such step sizes, the variability of m for concurrent series will be much like the variability of m for completely independent series. This matter deserves further checking, but it seems clear at the moment that the best opportunity for significant cross-correlational effects on the stability of m will occur when a small step size is chosen.

d. Implications for statistical tests of values of m. Until further information is forthcoming on the merits of various test procedures based upon s_m for short up-and-down series, we have two ways in which we may proceed in conducting significance tests concerning values of m. One is to use the

Table 8

EVALUATION OF s_m AS A PREDICTOR OF VARIABILITY IN THE VALUES OF m FOR CONCURRENT SERIES.

(Data for Group R-1, Experiment L-16, Part I)

Standard	No. of concurrent series	Computation of expected concurrent-series variability				Concurrent series data	
		Simple Av. s, in steps, for 8 subjects	Av. N_c per series	s_m, in steps	$s_{(m_1-m_2)}$, in steps	Root-mean-square concurrent-series diff. in m	No. of observed differences, 1 per subject
53 db.	2	1.91	24.3	.36	.51	.36	8
77 db.	2	.72	24.4	.15	.22	.19	8

formulae and procedures based on s_m, appreciating that our evaluation of differences in m may be conservative. The other is to use concurrent-series variability in the values of m as our error measure for m. Thus, for example, in a single-session experiment where two (or more) concurrent series are run under each of several test conditions, differences between conditions may be evaluated by analysis of variance procedures applied to the obtained values of m, disregarding completely the variability of the individual series. The latter procedure has been used and discussed in detail by McDiarmid (1962a).

15. SOME UP-AND-DOWN DATA ON THE PROBLEM OF SESSION-TO-SESSION VARIABILITY OF MEASURES ON A SINGLE SUBJECT: THE VARIABILITY OF $\bar{m}$ FOR DIFFERENT SESSIONS.

Preceding sections have discussed two questions related to the reliability of estimates of μ obtained by the up-and-down method: the variability of m across replications with the one-trial-per-subject design, and the variability of m for concurrent series with the same subject. Still another aspect of the reliability problem is that which concerns the variability of estimates of μ obtained during different experimental sessions with the same subject. Let us assume that the testing program for each session involves the use of concurrent series. We compute m for each up-and-down series, and find the average value of m (i.e. $\bar{m}$) for each session. How consistent are the values of $\bar{m}$ for different sessions? How is the variability of $\bar{m}$ related to available measures of within-session variability?

a. Source of data. For data on these questions we have examined the records of two experiments which will be discussed in more detail in a section soon to follow. The experiments were not conducted specifically for purposes of looking at session-to-session variability, but each subject in each of the studies was run twice under comparable conditions using the up-and-down method with three concurrent series. In the one study on duration discrimination, up-and-down testing constituted the entire experimental session on two of four test days. In the other study on stereoscopic discrimination, up-and-down testing constituted the opening half of two experimental sessions. So for all subjects we had, and report here in Table 9, information on the variability of $\bar{m}$ from one session to a second, where conditions were the same in both sessions.

b. Analysis. The nature of our analysis becomes clear in terms of our entries in Table 9:

Column 2 lists the average number of trials, $\bar{n}_c$, on which were based the computations of m and s for the individual series.

Columns 3 and 4 present the average value of s for the six up-and-down series for each subject, and the reciprocal of this which indicates the average step size for him in standard deviation units.

Column 5 gives the variance of m as estimated from $\bar{s}$.

Column 6 presents a measure of variability introduced here for the first time in these discussions and designated (by us) by the symbol, S. S was computed using the formula for s but taking as data the combined frequency distributions for three concurrent series. Thus we obtained but one value of S per session, representing a mixture of within-series and between-series variability for that session. $\bar{S}$ is the average value of S over the two sessions.

Column 7 lists the variance of $\bar{m}$ as estimated from $\bar{S}$, and represented by the symbol $S^2_{\bar{m}}$. Since each value of $\bar{m}$ entails three times as many observations as each value of m, $S^2_{\bar{m}}$ may be as small as, but cannot be smaller than, one-third the value of s^2_m.

Columns 8 and 9 provide the mean squares for concurrent series and for sessions obtained from an analysis of variance of the six values of m for each subject.

Finally, columns 10 through 13 give four variance ratios which are of interest, with those values which are "significant" at the .05 level indicated by asterisks. Although s^2_m is not a traditional variance measure, tests equivalent to those in columns 10 and 12 have been suggested as suitable (Anon., 1944).

Looking first at column 10, we see that for 12 of the 20 subjects, m for concurrent series was less variable than expected from s_m. This trend was due entirely to differences for those subjects with the smaller step sizes, and is in line with the discussion on page 47 above, to the effect that the likelihood of s_m overestimating the variability of m varies with step size. Unexpected was the finding that for two subjects there was "significantly" greater variability between values of m for concurrent series than expected from s_m. These two cases can only be ascribed to sampling error.

With regard to the variability of $\bar{m}$ from session to session, we see high variability for 4 of the duration subjects and 5 of the stereo subjects. The test in column 11 based on the analysis of variance is of relatively

Table 9

SUMMARY OF DATA ON SESSION-TO-SESSION VARIABILITY OF INDIVIDUAL SUBJECTS
(For description of column entries, see text)

-1-	-2-	-3-	-4-	-5-	-6-	-7-	-8-	-9-	-10-	-11-	-12-	-13-
							Mean Squares		Variance Ratios			
Subject	$\bar{n}_c$	$\bar{s}$, in steps	Av step size	s_m^2 in steps^2	$\bar{S}$, in steps	$S_{\bar{m}}^2$, in steps^2	based on m's for concurr. series: MS_{conc} in steps^2	based on $\bar{m}$'s for different sessions: $MS_{sess.}$ in steps^2	$\frac{MS_{conc}}{s_m^2}$	$\frac{MS_{sess}}{MS_{conc}}$	$\frac{MS_{sess}}{s_m^2}$	$\frac{MS_{sess}/3}{S_{\bar{m}}^2}$
Duration Study									df:4,84 $F_{.95}$=2.52	df:1,4 $F_{.95}$=7.71	df:1,84 $F_{.95}$=4.00	df:1,84 $F_{.95}$=4.00
G	14.5	1.42	.7$\bar{s}$	.130	1.50	.048	.082	.813	.63	9.88*	6.25*	5.65*
H	13.5	1.00	1.0$\bar{s}$	.074	1.28	.038	.224	.000	3.02*	.00	.00	.00
M	14.5	1.03	1.0$\bar{s}$	.073	1.06	.025	.046	.070	.63	1.51	.96	.93
V	14.5	.95	1.0$\bar{s}$	.063	1.11	.027	.140	.984	2.22	7.00	15.6 *	12.0 *
P	13.8	1.48	.7$\bar{s}$	.146	1.62	.057	.140	.047	.96	.33	.32	.17
S	14.7	.82	1.2$\bar{s}$	.049	.86	.018	.032	.004	.65	.12	.08	.07
G	14.5	1.03	1.0$\bar{s}$	.073	1.10	.027	.067	.052	.92	.78	.71	.64
H	14.7	.83	1.2$\bar{s}$	.0515	.93	.020	.087	1.612	1.69	18.6 *	31.3 *	26.9 *
B	14.3	.52	1.9$\bar{s}$	.0243	.55	.009	.034	.052	1.40	1.56	2.24	1.97
H	14.7	.73	1.4$\bar{s}$	.040	.79	.015	.044	.016	1.10	.36	.40	.35
F	14.7	1.25	.8$\bar{s}$	.101	1.33	.038	.079	.096	.78	1.21	.95	.84
C	14.2	.84	1.2$\bar{s}$	.052	.85	.018	.019	.992	.36	53.50*	19.1 *	18.3 *
Stereo Study									df:4,120 $F_{.95}$=2.45	df:1,4 $F_{.95}$=7.71	df:1,120 $F_{.95}$=3.92	df:1,120 $F_{.95}$=3.92
A	19.2	2.19	.5$\bar{s}$	.218	2.24	.075	.086	1.815	.39	21.2 *	8.3 *	8.1 *
G	19.5	1.24	.8$\bar{s}$	.075	1.30	.026	.040	.056	.53	1.4	.75	.72
R	19.3	1.86	.5$\bar{s}$	.160	1.92	.054	.093	.032	.58	0.3	.20	.19
K	19.3	1.86	.5$\bar{s}$	.160	2.24	.072	.270	.003	1.69	0.01	.02	.01
F	19.3	.87	1.2$\bar{s}$	.041	1.04	.018	.009	.184	.22	20.0 *	4.5 *	3.4
H	19.3	1.42	.7$\bar{s}$	.097	1.46	.034	.072	1.382	.74	19.1 *	14.2 *	13.6 *
J	18.8	1.26	.8$\bar{s}$	.080	1.38	.030	.112	7.216	1.40	64.7 *	90.2 *	80.2 *
P	19.7	1.02	1.0$\bar{s}$	.053	1.22	.024	.194	1.470	3.66*	7.6	27.7 *	20.4 *

low sensitivity because of the small number of degrees of freedom involved in the error term, but in general we have agreement with the tests given columns 12 and 13.

c. Discussion. Of interest is the extremely large size of many of the variance ratios for the session effect. These values imply very marked changes in μ between sessions. We may conclude that, at least for some subjects, marked drifts in μ or changes in the criterion of judgment occur from session to session. Whether such drifts are characteristic of all or most subjects will become more apparent when more critical experiments are run extending over more than two sessions. For such experiments, it appears that any of the three different variance ratios used in Table 9 to test session effects should be adequate.

In general the psychophysical literature has not given much attention to session-to-session change in the PSE. That such change might occur is of course not unexpected in terms of the occurrence of drifts within sessions. Perhaps the relative ease of evaluating between-session effects by the up-and-down method will lead to further investigation of this session problem.

16. AN EXPERIMENTAL COMPARISON OF THE UP-AND-DOWN METHOD AND THE METHOD OF CONSTANT STIMULI: RESULTS OF TWO EXPERIMENTS.

(a) Background. In the introduction to this report it was pointed out that the method of constant stimuli is a method in which session-wide stimulus context is controlled by the experimenter, and in which the judgment distribution or judgment context is a function of the subject's responses to those stimuli. In the up-and-down method, on the other hand, session-wide judgment context is controlled by the nature of the sequential program of trials, and the stimulus distribution or stimulus context is a function of the subject's discriminative behavior. Both methods are used to estimate μ and σ of the psychometric function. From the estimate of μ, we quantify the subject's bias, as (PSE-POE) or (m-POE), and in the estimate of σ we have a direct measure of the subject's differential sensitivity.

One of the early tasks of our project was to compare estimates of μ and σ obtained by these two methods. This comparison was motivated by the expectation, shared with other experimenters, that when a moderate to large judgment bias exists, attempts to measure it by the method of constant stimuli will typically result in measures which are too small. This expectation is based on the premise that the PSE obtained by this method will be constrained near the POE when the comparison stimuli are symmetrically distributed about the POE. There are two arguments for this view: (a) The PSE tends toward the center of the comparison series when the latter is asymmetrically located with reference to the POE (see, for example, Harris, 1948). Surely if this is so, it is reasonable to suppose that the PSE is also influenced by stimulus context when a symmetrically located series of comparison stimuli is used. This would keep the PSE near the POE, and would mean a depressed measure of bias. (b) Subjects in psychophysical experiments frequently appear to be set on disposed to use the two opposed responses, "heavier" vs.

"lighter", "louder" vs. "softer", etc., equally often. Such a set favors the PSE being near the POE if the comparison stimuli fall symmetrically about the POE.

The up-and-down method, by contrast, hunts for the PSE. It cannot constrain the PSE because it imposes no stimulus context upon the subject. Rather it allows the subject to have "whatever comparison stimuli he wants" in order to locate the PSE. And a second feature of the up-and-down method which is equally interesting is that while it does impose a judgment context on the subject, that judgment context is the 50-50 one which he appears to expect anyway.

These considerations lead to the hypothesis that a subject who has a judgment bias with regard to a given differential discrimination, will evidence a larger bias when tested by the up-and-down method than when tested by the method of constant stimuli. It may also be conjectured that such constraint as the method of constant stimuli may impose on the PSE will make PSE's determined by this method more uniform or stable from day to day or from subject to subject than PSE's determined by the up-and-down method.

b. Experiment CS-1: A comparison of the method of constant stimuli and the up-and-down method in a study of the discrimination of auditory durations.

(1) Specific purpose: Previous work by Stott (1935), Woodrow and Stott (1936), and others has shown that the time error for very short duration is positive, while that for longer durations is negative. The "indifference" duration, or duration where the transition occurs between positive time errors and negative time errors, is estimated to be in the range of 1 to 2 seconds. From the arguments presented above, we would expect the up-and-down method to reveal a larger positive time error than the method of constant stimuli when the duration being judged is short, say of the order of 0.5 seconds, and a larger negative time error when the duration being judged is long, say of the order of 5.0 seconds. In other words, the function relating time error to duration should be clearer or steeper for the up-and-down method than for the method of constant stimuli. Further, if the method of constant stimuli does place constraints upon the PSE in terms of context effects, then differences between the PSE's for different individuals should be less by the method of constant stimuli than by the up-and-down method.

Our purpose then was to compare the two methods with regard to time error magnitudes and individual differences in time errors.

(2) Apparatus and general procedure. The auditory stimulus used in this experiment was white noise. It was of moderate intensity, keyed by an electronic switch with 50 msec. rise time, and presented by loudspeaker located some six feet from the subject. The subject sat alone in a small experimental room and indicated his judgments by throwing one of two spring-loaded lever-type switches. He pushed one key if he judged the second sound to be longer in duration than the first, the other key if he judged it to be shorter than the first.

There were 31 stimulus durations available, each differing from its neighbors in the series by 1/24th of a log unit. The series ran: .38 secs., .41, .45, .50, .55, .61, .67, ...1.19, 1.31, 1.44, 1.58, 1.74, 1.92, ...3.75, 4.13, 4.54, 5.00, 5.50, 6.06, and 6.67 secs.

Trials for both the method of constant stimuli and the up-and-down method were presented using the automatic programmer described in the appendix to this report. This equipment was located in a room adjacent to the subject's room. Both the stimulus sequence and the subject's responses were recorded on a multi-channel unit using electro-sensitive recording paper.

The subject was alerted for each trial by a small panel light which came on as a warning signal 2 seconds before the standard. The standard was a sound of either .50, 1.58 or 5.00 seconds duration, the three values underlined in the above series. The inter-stimulus interval, following the standard and preceding the comparison, was always 5.00 seconds. The interval between trials, from the end of the comparison to the next warning signal was 9.75 seconds. Maximum times per trial were thus of the order of 25 seconds.

The subject made 90 judgments per day on four consecutive experimental days. After every 15 trials he had a two-minute rest. All sessions were completed in less than an hour. First day sessions were the longest in that they included instructions and a brief series of practice trials.

(3) Subjects. The subjects were 12 young men, either high school seniors or college undergraduates attending the 1959 Summer Session. All were paid for participating in the experiment, most in fact having been recruited through the Student Employment Service.

(4) Design. Each subject was tested with but one standard, making the design a random-groups design with four subjects per group. Two of the subjects in each group were tested using the method of constant stimuli on days 1 and 3, the up-and-down method on days 2 and 4, while the other

subjects were tested using the methods in the opposite order.

(5) Comparison series for the method of constant stimuli. Five comparison durations were used with each standard. One of these was equal to the standard. The others were the two durations immediately shorter than and the two immediately longer than the standard. Thus each comparison series was symmetrically distributed about the standard on the logarithmic scale of duration. The order in which the several comparison stimuli were used on successive trials was random, subject to the two restrictions that the same comparison never be used on two successive trials, and that each comparison occur three times in each set of 15 trials.

(6) Programming for the up-and-down method. Three concurrent series with the same standard were used in the up-and-down sessions. Each series was programmed using a different add-and-subtract stepper. The starting level for each series was randomly chosen--either equal to the standard, the standard minus one level, or the standard plus one level. The sequence in which trials were taken from the three series was based on random permutations of the three, subject to the restriction that no given series be used on two successive trials. On the second experimental day with the up-and-down method, each series was continued where it had left off at the end of the first day.

(7) Determination of the PSE's and difference thresholds. Data for each session under the method of constant stimuli were tallied in the usual way, plotted on normal probability paper and a straight line fitted to the points by eye. The latter operation was done independently by two experimenters. The plotted probabilities were based upon 18 observations per comparison stimulus for each experimental session. From each fitted line, the value of the median and the standard deviation were read as estimates of μ and σ of the psychometric function. The median and standard deviation for each session are hereafter designated as Mdn_{VL} and SD_{VL}, where the subscript stands for "visual line." Their values are averages of the two experimenters' estimates.

For the up-and-down method, m and s were computed for each series separately on each experimental day. The values of m for the three series were averaged to obtain $\bar{m}$ for the session. Values of s were averaged to obtain $\bar{s}$. All calculations were based on all trials of the up-and-down series, in as much as m was never very far from the initial testing level and biasing effects of that testing level must have been very small.

As will be explained below, the data for all up-and-down sessions were also subjected to graphic analysis -- partly as a check that the outcome of the experiment did not hinge on differences in analysis between the up-and-down and constant methods, and partly as a check on the degree of agreement of this analysis with that employing the Princeton formulae.

The results of the study are summarized in Figures 5 and 6 and in Table 10;

(8) Results on time error magnitudes. Figure 5 is a scatterplot of bias measures, subject by subject, by the two psychophysical methods. The data plotted are the two-session averages for each subject, as listed in columns 3 and 7 of Table 10. Had bias measures been the same or similar by the two methods, the points would have been fitted well by the 45-degree line in the figure. As will be seen, 3 of 12 subjects had essentially no bias as measured by either method (less than 0.1 stimulus step). For 1 of the remaining 9 subjects, a larger bias was found in data obtained by the method of constant stimuli, for 1 the bias was equal by both methods, and for 7 the measured bias was larger by the up-and-down method. It may also be noted that of 7 bias measures which were 0.5 stimulus step or larger, 6 were obtained with the up-and-down method.

(9) Results on time errors vs. duration. The relation of the time error measurements to stimulus duration is plotted in Figure 6. The fact that the function is steeper for data obtained by the up-and-down method is interpreted to mean that the function determined by the method of constant stimuli was flattened by stimulus context factors which limit bias measures by that method.

(10) Results on individual differences. The range of bias measures in each group of subjects was clearly greater for the up-and-down method. This can be seen in Figure 6, as well as in a comparison of the values listed in columns 3 and 7 of the table.

(11) Results on session-to-session differences. Session differences for the two methods are compared in columns 2 and 6 of the table. For 6 of the 12 subjects, session-to-session differences were larger by the method of constant stimuli and for 6 they were smaller.

Given the condition of fully independent observations, the up-and-down method is known to be more efficient than the method of constant stimuli, requiring some 30% fewer trials for the same precision of measurement (Dixon and Mood, 1948). Thus, for the same number of trials the standard error of m should be less than the standard error of mdn_{VL},

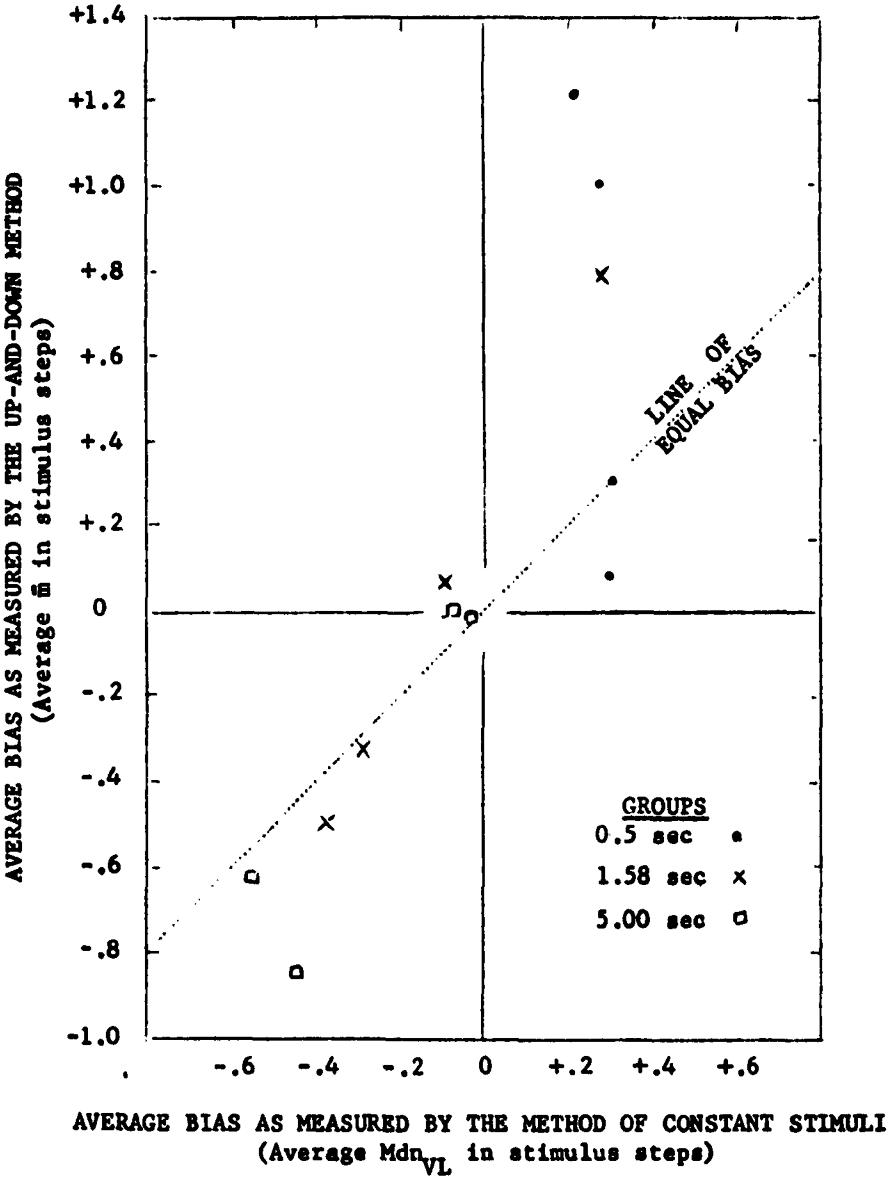

Figure 5: Relation between estimates of μ by the method of constant stimuli and by the up-and-down method in the duration study. The plotted points are for the 12 individual subjects.

Table 10

BIAS IN DIFFERENTIAL JUDGMENTS OF AUDITORY DURATION,
AS MEASURED BY THE UP-AND-DOWN METHOD AND BY THE METHOD OF CONSTANT STIMULI.

Note: In the present table values of $\bar{m}$ and mdn_{VL} are given directly as bias measures in stimulus steps. Estimates of σ are also given in stimulus steps.

		Data for the up-and-down method				Data for the method of constant stimuli			
Group with 0.50 sec. standard:		$\bar{m}$ for each session	session to session diff. in $\bar{m}$	Av. $\bar{m}$ over 2 sessions	Av. s over 2 sessions	mdn_{VL} for each session	session to session diff. in mdn_{VL}	Av. of mdn_{VL} over 2 sessions	Av. of ests. of σ over 2 sessions
Subj.	Day	-1-	-2-	-3-	-4-	-5-	-6-	-7-	-8-
G	1	+1.58	.74	+1.21	1.42	+0.45	.48	+0.21	1.15
	2	+ .84				+0.03			
H	1	+1.01	.01	+1.00	1.00	-0.08	.70	+0.27	1.58
	2	+1.00				+0.62			
M	1	+0.41	.22	+0.30	1.02	+0.22	.16	+0.30	1.31
	2	+0.19				+0.38			
V	1	+0.49	.81	+0.08	.95	+0.38	.18	+0.29	.80
	2	-0.32				+0.20			
Av.				+0.65				+0.27	
roup with .58 sec. tandard:									
P	1	+0.71	.17	+0.79	1.48	+0.52	.50	+0.27	1.29
	2	+0.88				+0.02			
S	1	+0.09	.05	+0.06	.81	0.00	.20	-0.10	1.06
	2	+0.04				-0.20			
G	1	-0.41	.19	-0.32	1.03	-0.50	.42	-0.29	1.04
	2	-0.22				-0.08			
H	1	-1.02	1.03	-0.50	.84	-0.45	.15	-0.38	.71
	2	+0.01				-0.30			
Av.				+0.03				-0.12	
roup with .00 sec. tandard:									
H	1	-0.06	.10	-0.01	.73	+0.20	.48	-0.04	.94
	2	+0.04				-0.28			
B	1	+0.09	.18	0.00	.52	-0.15	.15	-0.08	.60
	2	-0.09				0.00			
F	1	-0.50	.25	-0.62	1.25	-0.70	.28	-0.56	.91
	2	-0.75				-0.42			
C	1	-1.26	.81	-0.85	.85	-0.58	.26	-0.45	1.19
	2	-0.45				-0.32			
Av.				-0.37				-0.28	

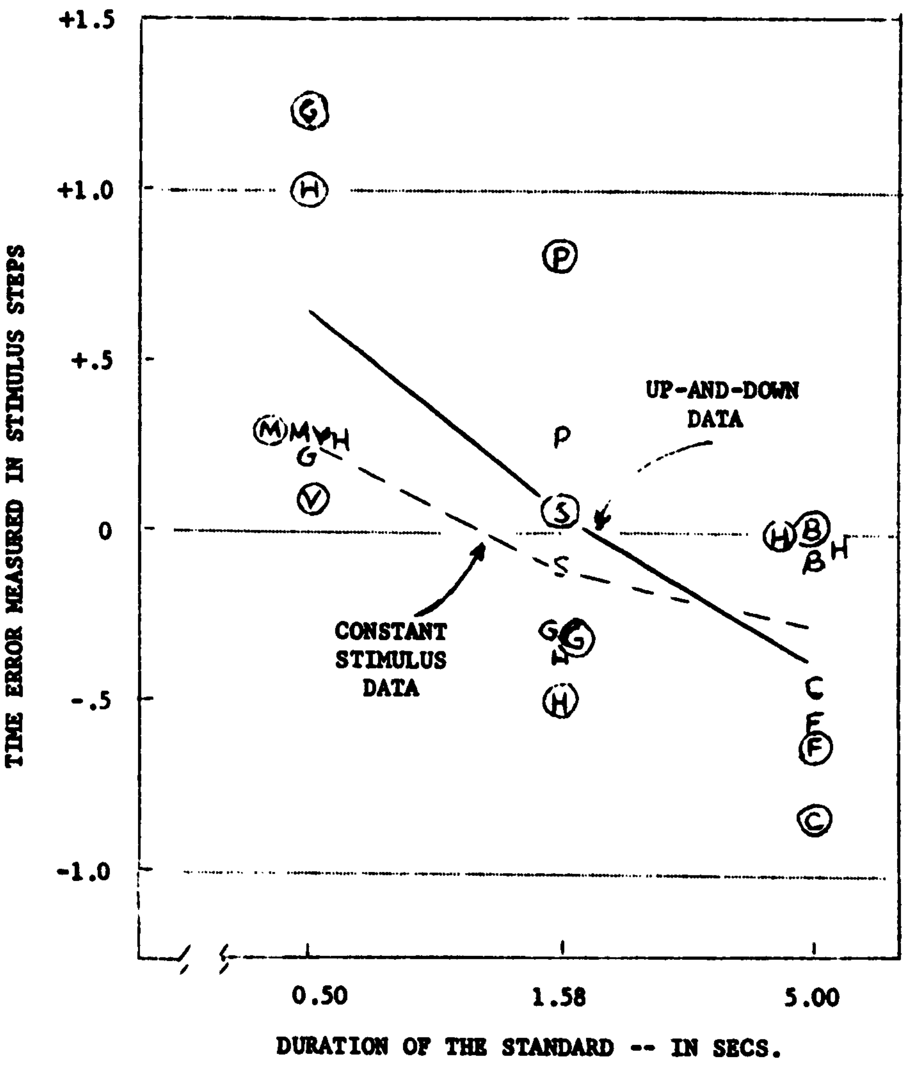

Figure 6: Time error for auditory duration as a function of the duration of the standard stimulus. Letters are subjects' initials. Circled symbols = up-and-down measures. Uncircled symbols = constant stimulus measures.

even if we forget that our random errors in visually estimating the line of best fit must also increase the standard error of mdn_{VL}. Opposing this trend and working to decrease variability in mdn_{VL}, however, is restriction of bias as a result of stimulus context. We had supposed that this effect might be strong enough that session-to-session differences in estimates of μ would be even less by the method of constant stimuli than by the up-and-down method. The data indicate no such extreme result, but bias restriction was apparently sufficient that the method of constant stimuli did not exceed the up-and-down method in session-to-session variability.

(Our interest here, of course, has been only in a comparison of session-to-session variabilities for our two methods. An evaluation of session-to-session variability of the single subject tested with concurrent series during each session has appeared above in section 15.)

(11) Summary. As had been expected, larger time errors were obtained in the discrimination of auditory durations when the up-and-down method was used than when the method was that of constant stimuli. Individual differences in judgment bias were also greater by the up-and-down method. These results support the view that PSE's obtained by the method of constant stimuli are constrained to be near the POE, making measures of bias improperly small by that method.

Concerned in this experiment were Drucker, McDiarmid and the author.

c. Analysis of up-and-down data for experiment CS-1 in terms of response proportions. It is of interest as a control for the foregoing results, and in a more general sense also, to examine the consequences of analyzing up-and-down data by the same procedures used in processing data collected by the method of constant stimuli. We have therefore compared the estimates m and s with parallel estimates of μ and σ obtained by applying graphic methods to response proportions calculated for the different comparison stimulus levels used in the up-and-down sessions.

For this analysis, all responses for the three concurrent up-and-down series for a given subject on a given day, were combined into a single distribution for that session. The proportion of "longer" judgments was computed for each stimulus level and plotted on normal probability paper. A visually determined line of "best" fit was drawn to the plotted points, and estimates of μ and σ read from the line. These were compared with $\bar{m}$

and S computed for that session (i.e. with average m and average s for the concurrent series). The record of 12 subjects, each tested on 2 days, provided 24 sessions on which comparative data were assembled.

Several features of an analysis of up-and-down data in terms of response proportions may be anticipated. These included the following:

(1) Whatever the number of levels used in the up-and-down series, obtained response proportions will typically be more reliable for stimulus levels near the center of the range of levels, for it is the primary feature of the up-and-down method that trials will be concentrated near the .50 point of the psychometric function. This suggests, in fact, that the two points bracketing the median might be taken as the principal basis for forming the line of best fit. In the extreme, one might merely interpolate between these two to locate the median, and this we have also done below.

(2) Whatever the number of levels used in the up-and-down series upon which the response proportions are based, the uppermost stimulus level has associated with it nothing but "longer" responses, while the lowest stimulus level has no "longer" responses. In general it appears that these extreme proportions of 1.00 and .00 should be set aside, not only because they cannot be located properly on the normal probability plot, but also because they are typically of low reliability. This makes the "line of best fit" a line which is fitted to two fewer proportions than the number of stimulus levels used.

(3) If only 4 levels have been used, there are but 2 points to plot. In the event that these do not bracket the median, the median is not clearly determined. Of our 24 series, 4 involved 4 levels and in one of these the points did not bracket the .50 point.

(4) If only 3 levels have been used (i.e. if stimulus step size was very large), the plot leads to no estimate of the psychometric function.

(5) If some larger number of levels has been used, say 6 or 7, the chances are quite good that some of the proportions near the extremes will be based upon very few observations and that their unreliability will introduce appreciable non-linearity of the plot on normal probability paper, even though the psychometric function is in fact normal. In our data, 3 of the 7 sessions where 6 or 7 levels were used resulted in plots which departed markedly from linearity.

(6) If 4 or 5 levels have been used, the individual proportions

are based on more trials and thus are more reliable than the proportions obtained when 6 or 7 levels are used (given, of course, that $N_{tot.}$ is the same in both cases). This should favor greater reliability of the graphically determined medians in the 4-5 level case, and presumably better agreement of these medians with the corresponding values of m determined from the up-and-down series by the Princeton formula. This is borne out in the data summarized in Table 11.

Table 11 presents, for each test session, the value of $\bar{m}$, the value of mdn_i (the median read by interpolating between the two proportions which spanned the .50 point), values of mdn_{VL} and SD_{VL} (median and standard deviation read from the visual line of best fit), and the value of $\bar{s}$.

From the table we may extract and compare the data for those 17 sessions in which 4 or 5 stimulus levels were used, and the data for the 7 sessions in which 6 or 7 levels were used. The former, quite obviously, were sessions where the values of $\bar{s}$ were lower (average value of $\bar{s}$ = .83 steps), while the latter were those of higher variability (average value of $\bar{s}$ = 1.38 steps). There were relatively more non-linear plots which presented line-fitting problems in the case of the 6 or 7 level sessions, as expected from (5) above (3 out of the 7 sessions as compared with 2 out of 17). The deviations of mdn_{VL} from $\bar{m}$ averaged 0.05 stimulus steps for 4-5 level sessions and 0.08 for 6-7 level sessions. To remove the effect of s on these measures, we took as our index of the degree of disparity between mdn_{VL} and $\bar{m}$ the absolute value of $(\bar{m} - mdn_{VL})/\bar{s}G$. This quantity was 20% smaller, on the average, when 4 or 5 levels had been used than when 6 or 7 levels had been used. Presumably this difference is associated with the greater reliability of the individual proportions in the 4-5 level case, as expected under (6) above.

For the most part, values of mdn_i were very similar to mdn_{VL}. Values of mdn_i deviated slightly more from $\bar{m}$ than did the values of mdn_{VL}, but it is clear that agreement among the three quantities was very good for the most part.

Values of SD_{VL}, it will be noted, were typically larger than corresponding values of $\bar{s}$. This is as it should be, considering the fact that SD_{VL} includes variance between concurrent series, which is not included in $\bar{s}$. SD_{VL} and S (see page 50) did not differ systematically.

What we have found then is that with 90 test trials per session and a step size where 4 or 5 stimulus levels are used, a "visual line of best fit"

Table 11

RESULTS FOR GRAPHIC METHODS OF ESTIMATING PARAMETERS OF THE PSYCHOMETRIC FUNCTION FROM UP-AND-DOWN DATA.

		Estimates of μ			Estimates of σ	
		Av. of the 3 estimates of μ by Princeton formula	Interpol. Median on prob. Paper	Estimate from visual line of good fit on prob. paper	Av. of the 3 estimates of σ by Princeton formula	Estimate from visual line of good fit on prob. paper
Group with 0.50 sec. standard						
Subj.	Day	$\bar{m}$	mdn_I	mdn_{VL}	$\bar{s}$	SD_{VL}
G	1	+1.58	+1.85	+1.80	2.14	1.55
	2	+ .84	+ .80	+ .85	.70	.70
H	1	+1.01	+1.05	+1.05	1.00	1.15
	2	+1.00	+ .95	+1.05	1.01	1.25
M	1	+ .41	+ .45	+ .45	1.22	1.20
	2	+ .19	+ .20	+ .20	.83	.90
V	1	+ .49	+ .40	+ .40	.82	1.05
	2	- .32	- .30	- .35	1.08	1.20
Group with 1.58 secs. standard						
Subj.	Day					
P	1	+ .71	+ .65	+ .80	1.99	2.25
	2	+ .88	+1.05	+ .95	.97	1.20
S	1	+ .09	+ .10	+ .05	.94	1.00
	2	+ .04	+ .10	+ .05	.69	.75
G	1	- .41	- .25	- .50	1.10	1.10
	2	- .22	- .25	- .30	.96	1.05
H	1	-1.02	-1.20	-1.10	1.10	1.05
	2	+ .01	- .05	+ .10	.57	.70
Group with 5.00 secs. standard						
Subj.	Day					
H	1	- .06			.57	
	2	+ .04	+ .10	+ .10	.89	1.00
B	1	+ .09	+ .05	+ .05	.56	.70
	2	- .09	- .05	- .05	.48	.55
F	1	- .50	- .45	- .50	1.37	1.25
	2	- .75	- .70	- .80	1.13	1.45
C	1	-1.26	-1.40	-1.30	1.33	1.45
	2	- .45	- .40	- .40	.36	.45

to up-and-down data provides estimates of μ and σ which on the average are very close to m and s. The mdn_{VL} differed from $\bar{\bar{m}}$ on the average by only 0.05 step in the present data, and the value of SD_{VL} differed from $\bar{\bar{s}}$ on the average by about 0.10 stimulus step. Clearly those differences reported above between bias measures by the up-and-down method and by the method of constant stimuli are not to be ascribed to peculiarities of the graphic method which we had used for treating the constant stimulus data.

d. Experiment CS-2: A comparison of the method of constant stimuli and the up-and-down method in a study of stereoscopic discrimination. This experiment by Kappauf and Arbit was similar in general objective to the duration study but followed a different plan. In this case the subject was tested on a preliminary experimental day by the up-and-down method to obtain an estimate of his PSE -- i.e. the position of a variable pin which made it appear equidistant with a fixed pin. On subsequent days, comparison stimulus conditions for the method of constant stimuli consisted of 5 pin positions which were symmetrically distributed about this preliminary PSE. Up-and-down measures were also taken on these days. The object was to learn whether these later estimates of μ would vary less from the preliminary one in the case of the method of constant stimuli than in the case of the up-and-down method.

This was a non-project experiment, supported in part by the Research Board of the University of Illinois, but is reported here because of its relation to duration experiment above.

(1) Apparatus and general procedure. The subject observed two pins from a distance of 5 meters and judged which of the pair was farther away. The pins were 1.5 mm. in diameter and were separated laterally by 6 cm. The subject viewed the central vertical segment of each of these pins through an aperture 3 cm. high and 22 cm. wide in a large white occluding screen behind which the experimenter stood. The pins were painted black, and through the aperture were seen against a white background panel which was 5.5 meters from the subject. Luminance levels of the occluding and the background screens were 5 and 7 foot lamberts respectively. The pins were exposed to view when the experimenter raised a sliding panel behind the opening.

The right hand pin was the movable or variable one. The subject was required to report the position of this pin as "nearer" or "farther" than the left hand one. He made this report when the sliding panel was lowered ending each 3-second exposure. He was cautioned against making head movements while viewing the pins. A new exposure or trial began every 10 seconds.

(2) Preliminary measures. On the preliminary day, the subject made 300 observations, with a rest period allowed after every 50 (i.e. about every 8 minutes). Observations were taken using 3 concurrent up-and-down series and a stimulus step size of 1 cm. Values of $\bar{m}$ and $\bar{s}$ were computed for each subject.

(3) Test sessions. Four test days followed the preliminary day. On each of these, each subject made 120 observations by the method of constant stimuli and 120 observations programmed by the up-and-down method using three concurrent series. Method order was counterbalanced on successive days for the the same subject and for any given day was balanced across subjects. Observations during these test days were conducted under conditions which depended upon $\bar{m}$ and $\bar{s}$ for the preliminary day. Those subjects for whom $\bar{s}$ was between 0.7 and 2.0 stimulus steps were tested for the remainder of the experiment using a step size of 1 cm. Those whose standard deviations were between 2.0 and 5.0 steps were tested on remaining days using a 2 cm. stimulus step. Two subjects were sufficiently precise in their preliminary judgments that a stimulus step of 0.5 cm. was chosen for them.

Each subject was tested with a set of comparison positions of the right hand pin which included one position at $\bar{m}$ from the preliminary run. Other comparison positions ranged forward and back from $\bar{m}$ by steps of the chosen size. For testing by the method of constant stimuli, the five chosen positions were symmetrically distributed about the preliminary $\bar{m}$. For up-and-down testing, many comparison positions might be used.

(4) Subjects. The subjects included the two experimenters and 6 others who were students.

(5) Analysis of the data. The analysis proceeded in the same manner as for the duration experiment. For the method of constant stimuli, the proportion of "farther" judgments obtained at each comparison pin position on each experimental day was plotted on normal probability paper and a line of best fit adjusted to the plotted points by eye. This was done independently by two judges, and the mean of their estimates of μ and σ was determined. This provided a value of mdn_{VL} and of SD_{VL} for each subject on each of four experimental days. Up-and-down data for each session were

processed to provide a daily value of $\bar{m}$ and $\bar{s}$. The up-and-down judgments were also assembled into a composite daily distribution and plotted as response proportions to which a line was fitted on normal probability paper.

(6) Results. Our results are summarized in Table 12 and in Figure 7.

For our basic comparison of the methods of data collection, we again compared mdn_{VL} and SD_{VL} for the constant stimulus data with $\bar{m}$ and $\bar{s}$ for the up-and-down data. For every one of the 8 subjects, the difference between $\bar{m}_{pre.}$ and the average value of $\bar{m}$ for the four day test period was greater than the difference between $\bar{m}_{pre.}$ and the average value of mdn_{VL} for the four day period. Some of the differences for individual subjects were small, but the scatterplot of the results in Figure 7 is very much like that for the duration study. Thus we must conclude again that the observed PSE on any given day was contrained to be near the middle of the comparison series when the method of constant stimuli was used.

Variability of the daily values of $\bar{m}$, as measured by the range of these values, was greater than the range of the daily values of mdn_{VL} for 6 or the 8 subjects. Again, as in the duration study, we fail to find measures by the up-and-down method more consistent as expected from "reliability" considerations. Rather, we find relatively less variability of the estimates of μ by the method of constant stimuli, implying constraint on the daily values of mdn_{VL}.

With regard to estimates of σ, the method of constant stimuli gave larger estimates for 5 subjects, the up-and-down method for 3. These results, with the duration data in Table 10 (p. 58), imply that the methods produce equivalent measures of differential sensitivity as measured by s or SD.

Graphic exaimination of response proportions computed from up-and-down records provided results in complete accord with those in the duration study. The mdn_{VL} for up-and-down sessions agreed with $\bar{m}$ within 0.05 stimulus steps on the average when the number of levels used was 4 or 5. The disparity between these two measures advanced to 0.10 stimulus steps on the average when the number of levels used was 6 or 7.

e. Summary comparison of the methods. When the up-and-down method and the method of constant stimuli are each used for continued testing with a given subject, the method of constant stimuli provides PSE's which are more stable and which depart less from the POE than does the up-and-down method. In an operational sense the up-and-down method is the less reliable,

Table 12

DATA FOR STEREOSCOPIC JUDGMENT,
OBTAINED BY THE UP-AND-DOWN METHOD AND BY THE METHOD OF CONSTANT STIMULI

Note: As in Table 10, values of $\bar{m}$ and mdn_{VL} are given in stimulus steps. Here they are given as deviations from the preliminary $\bar{m}$ described in the text.

Subject	Size of stimulus step	Average value of:		Range of values of:		Range of values of:	
		$\bar{m}$ over 4 days	mdn_{VL} over 4 days	$\bar{m}$ over 4 days	mdn_{VL} over 4 days	$\bar{m}$ over 2 days when up-down trials were first	mdn_{VL} over 2 days when const. stim. trials were first
A	2 cm.	+1.83	+0.91	1.11	.47	1.11	.30
G	2 cm.	-1.26	-1.18	.73	.88	.16	.87
R	1 cm.	-0.10	-0.05	.45	.40	.15	.32
K	0.5 cm.	+0.39	+0.21	.76	.54	.14	.47
F	1 cm.	+0.08	+0.02	.49	.63	.38	.35
H	1 cm.	+1.44	+0.22	.99	.15	.99	.15
J	1 cm.	-1.09	-0.60	2.19	1.03	2.19	.52
K	0.5 cm.	+0.38	+0.22	1.35	.20	1.00	.10

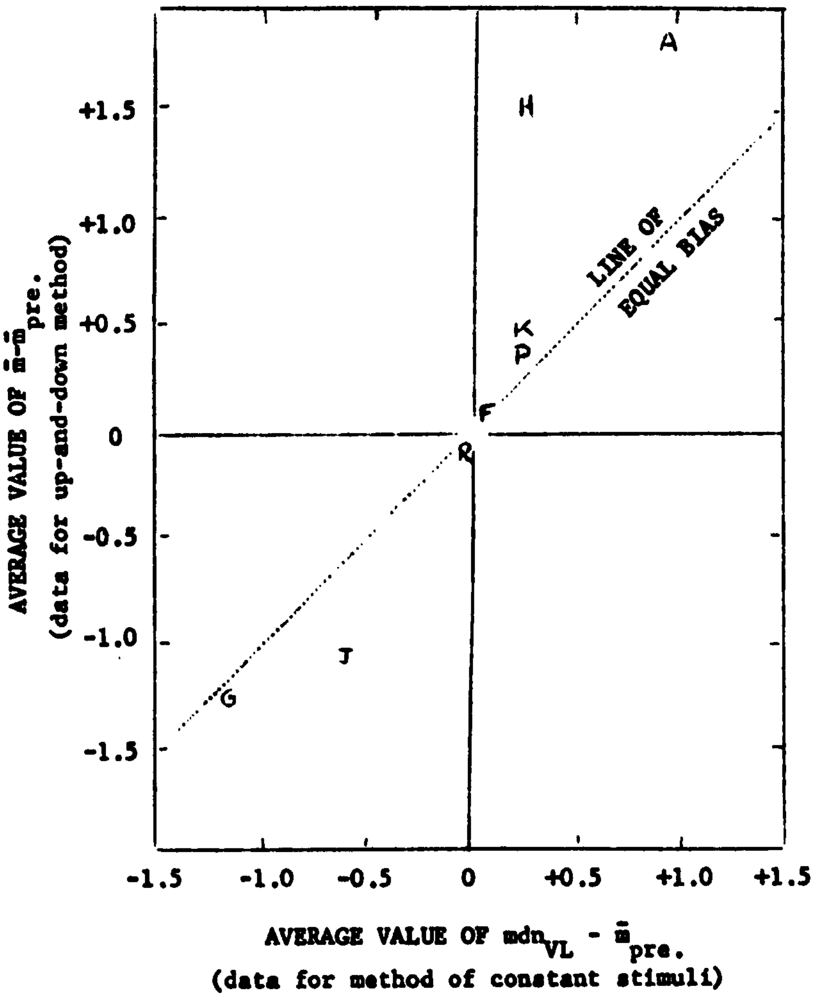

Figure 7: Relation between estimates of μ by the method of constant stimuli and by the up-and-down method in the study of stereoscopic judgments. The plotted points are for the individual subjects. Both axes are scaled in stimulus steps.

although theoretically it is the more reliable. The foregoing experiments are interpreted to mean that results for the method of constant stimuli are of spuriously high consistency because of the operation of stimulus context factors associated in the fixed set of comparison stimuli.

It was noted above that Fernberger (1913) had devised the scheme of embedding trials by the method of limits in a program of trials by the method of constant stimuli in order to conceal the sequential character of the method of limits series. Suppose now we make a full turn about on this scheme and embed trials by the method of constant stimuli in a program of trials by the up-and-down method (perhaps two concurrent series). Our intent would be to "conceal" the stimulus context provided by the comparison stimuli on constant stimulus trials. We would accomplish this by merging this context with the stimulus context of the up-and-down trials. In other words, both up-and-down and constant stimulus methods would share the same stimulus context. We expect that the estimates of μ should then vary for the two methods in the direction indicated by theory. This experiment has not yet been run, but deserves early attention.

17. A COMPARISON OF THE UP-AND-DOWN METHOD AND THE METHOD OF LIMITS.

The method of limits, in its traditional form, employs a sequential scheme of stimulus presentation and in this respect is a member of the same family of measurement methods as the up-and-down method. In the method of limits, conditions are changed in an orderly way from trial to trial on the basis of the subject's responses. Every response X determines that conditions for the next trial will move one step along the stimulus scale in the direction of being more favorable to the occurrence of response Y. The occurrence of a Y response (or some prescribed number of Y responses) identifies the "limit" for that series of trials, whereupon the experimenter may begin a new series by jumping back to a stimulus level where X is again almost certain to occur. The new series of trials proceeds as did the first. In this form, the method of limits could be described as "a modification of the up-and-down method" with stimulus steps in one direction many times larger than the steps taken in the other direction. Oppositely, the up-and-down method is likened by many to a "continuous" or "progressive" form of the method of limits, and we find Guilford (1954) considering the up-and-down method as one of the variations of the method of limits. More specifically, we may say that the similarity of the two methods in so far as their data collection operations are concerned resides in the fact that their programs of trials both follow Markov designs (see Smith, 1961).

It should be apparent, then, that the method of limits and the up-and-down method must be closely related in certain of their quantitative features. Interestingly, although the method of limits has seen considerable use by psychologists, there has been relatively little discussion of its quantitative properties. We take this opportunity therefore to consider some of these properties in a direct comparison of the method of limits with the up-and-down method.

a. Computation of the expected distribution of trials and the expected distribution of limits. The computation of these expected distributions proceeds in similar fashion for both methods, based simply upon the response probabilities at each stimulus level and upon the assumption of trial-tö-trial independence.

Urban (1907, 1908) appears to have been the first to reason from response probabilities and indicate the manner in which one may derive the expected distribution of limits under the method of limits. Suppose that we have a series of response probabilities such that at stimulus level i, q_i is the probability of response X and p_i is the probability of response Y. For convenience here we assume that q_1 is so close to 1.00 that we may take its value to be 1.00. Then the following computations apply:

Stimulus Level	Prob. of response X	Prob. of response Y	Prob. of last X occurring at this level.	Prob. of first Y occurring at this level.
1	$q_1 = 1.00$	$p_1 = .00$	q_1p_2	$p_1 = .00$
2	q_2	p_2	$q_1q_2p_3$	q_1p_2
3	q_3	p_3	$q_1q_2q_3p_4$	$q_1q_2p_3$ etc.

In the last column we have the probability distribution of limits, i.e. the expected distribution of limits, under the condition that we define the limit for each trial series as that stimulus level at which the first Y response occurs. In the next to last column we have the expected distribution of limits if we define the limit for each series of trials as that stimulus level at which the last X response occurs. Clearly these two distributions are identical except for the fact that one if offset from the other by one stimulus level.

Working back from the first of these distributions of limits we may quickly establish the distribution of trials which would be accumulated if many series of trials were conducted.

b. Dependence of the expected distributions of trials and of limits upon stimulus step size. Just as the expected distribution of trials depends on step size in the up-and-down method, so the expected distributions of trials and of limits depend on step size in the method of limits.

The first to recognize the influence of step size on the limits obtained in individual series of observations, and hence on the average limit obtained for a number of series, appears again to have been Urban (1907, 1908). Urban noted that it was common practice for experimenters to prefer small step sizes because these presumably would assure greater precision in measuring the limit. His probability analysis, however, led him to observe that small step sizes shift the expected value of the limit in the direction of "anticipation." His summary comment was: "... the result of a determination of the threshold by the method of just preceptible differences depends somewhat on the size of the intervals which are used. It is therefore not necessarily a sign of incomplete training of the subject or of his inability to direct his attention to the comparison of the stimuli, if series with small differences (i.e. stimulus steps) fail to give the same results as series with large differences." (1908, p. 60). Later, Fernberger (1913) working under Urban's direction performed a weight lifting experiment in which he systematically varied step size over a five-fold range, all step values being below 1σ. Fernberger did not comment on the matter, but it is clear from his data (his Tables XLII and XLIII) that "errors of anticipation" increased as the stimulus steps grew smaller in the manner expected from Urban's analysis.

The complete picture, here, can be developed if we assume some specific form for the psychometric function, and carry out Urban's calculations from the insert table above for a variety of step sizes. When we perform such an analysis, using step sizes ranging from 0.1σ to 3.0σ and assuming the psychometric function to be the cumulative normal, the results which we obtain are those summarized in Figure 8.

Each of the values plotted in the figure is the mean of the expected distribution of limits for some particular condition, i.e., the "expected"

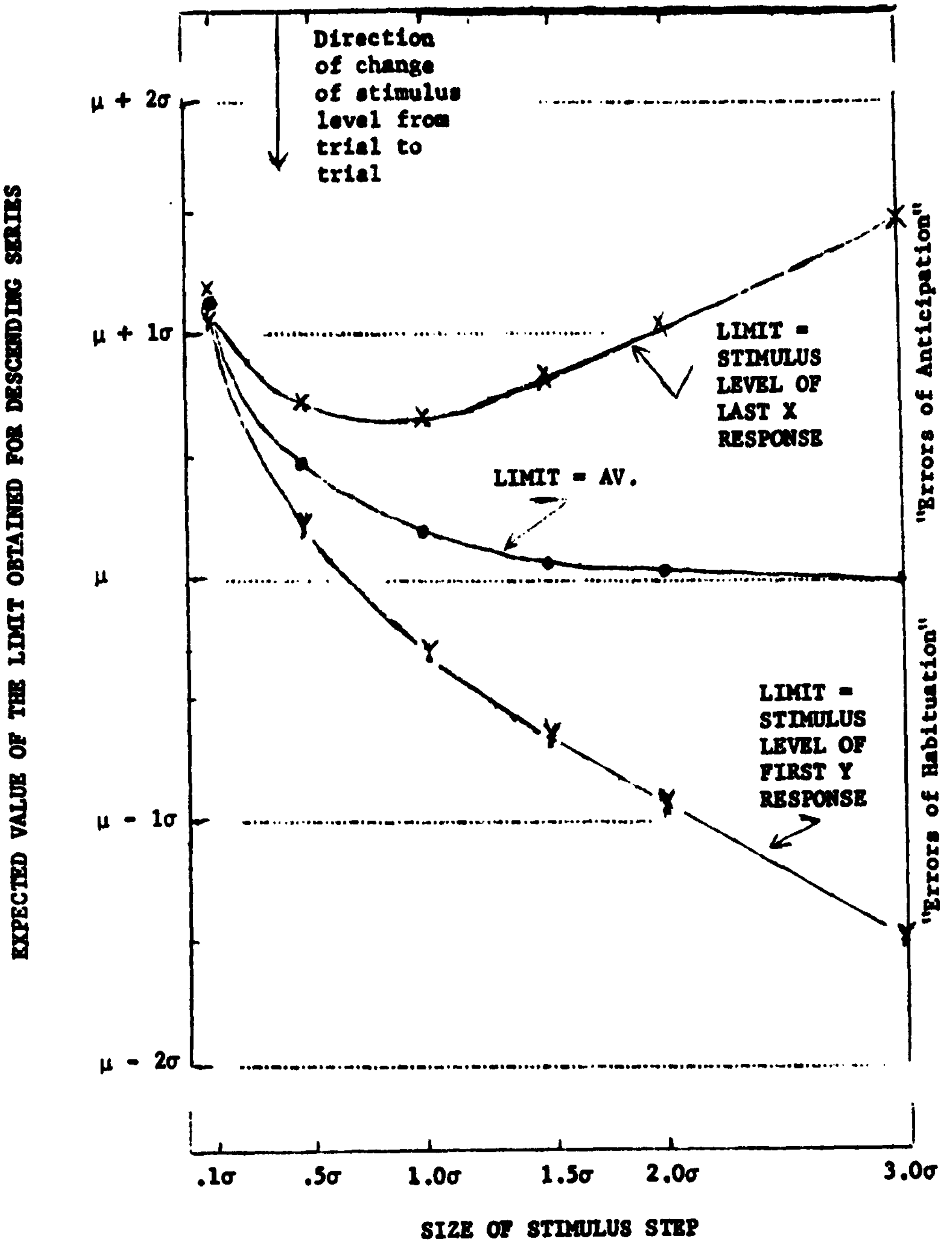

Figure 8. The effect of size of stimulus step on the expected value of the limit for descending series. Values of the limit are expressed relative to μ, the mean-median of the normal psychometric function. Limits have been computed under three definitions: (a) as the stimulus level where the last X response occurs; (b) as the stimulus level where the first Y response occurs; and (c) as the average of the stimulus levels of the last X response and first Y response.

limit for that condition. We have considered limits defined in three different ways: that the limit be (1) the stimulus level of the last X response, (2) the stimulus level of the first Y response, or (3) the average of these two (i.e., the stimulus level midway between that of the last X and the first Y response). Note that the figure deals with limits obtained for "decreasing" or "descending" stimulus series only. Symmetrical functions would apply for "ascending" series.

As we see from the figure, if the limit is defined as the stimulus level where the last X response occurs, the expected limit consistently occurs before the mean-median of the psychometric function is reached, i.e. it deviates in the direction described in the literature as "anticipation." If the limit is defined oppositely as the stimulus level where the first Y response occurs, the measure is biased in the direction of anticipation for small step sizes but shifts to become biased in the direction of habituation for step sizes larger than about 0.7σ. Lastly, if the limit is defined as the stimulus level midway between that where the last X and that where the first Y occurs, the measure is biased consistently in the direction of anticipation, but this bias becomes smaller and smaller as step size becomes larger.

We see then that, under our third and most commonly used definition of the limit, an error of anticipation is a statistical property of the method of limits. This renders meaningless many of the discussions of "errors of anticipation" which have appeared in the literature ascribing such errors to the subject. And this was Urban's point, of course.

An interesting property of the functions plotted in Figure 8 is that they are essentially independent of the position of the mean-median of the psychometric function relative to the testing levels: expected limits when the mean-median is midway between testing levels and when it coincides with some testing level, agree within the limits of graphing accuracy.

Recent discussions of this bias in the "one-way" method of limits appear in Anderson, et al (1946), McCarthy (1949) and Brown and Cane (1959). A scheme for adjusting one-way measures in terms of estimated step size is given by Anderson et al, but two-way measures are clearly preferred.

c. The form of the expected distribution of trials. The up-and-down method is clearly designed to concentrate trials in the vicinity of the mean-median of the psychometric function, and how well it does this we have already seen (page 15). For a one-way method of limits, i.e. where approaches to the mean-median are made from one direction only, the method obviously concentrates trials at that one end of the stimulus scale. But if both ascending and descending series are run, as is usually the case, then the method of limits provides some concentration of trials near the mean-median when the step size is not too small. Thus when step size is between 1σ and 2σ, the range of interest in earlier discussions, more trials (by a factor which is admittedly not large) occur at levels near the mean-median than at each level at the extremes of the stimulus scale. In general, however, the distribution of trials by the method of limits resembles the rectangular distribution of the method of constant stimuli much more than it does the peaked distribution of trials obtained with the up-and-down method.

d. Bias as a function of initial testing level. It was noted that initial testing level introduces bias in the estimate of μ by the up-and-down method if that level is far from μ and if step size is small. In the case of the method of limits, a similar source of bias exists -- bias to be added algebraically to that shown in Figure 8. Computations for Figure 8 assumed that each stimulus series would begin far enough from μ that the probability of response X was exceedingly close to 1.00. If the initial testing level is at a level where the probability of response X is not close to 1.00, the expected limit will deviate less in the direction of anticipation than shown in the figure (See Anderson, et al, 1946, p. 106).

It is interesting to note in this connection, that in Fernberger's methodological comparison of the method of limits and the method of constant stimuli (1913) he computed expected limits for ascending and descending series and compared their average with the mean-median of the psychometric function. These values should have agreed in Fernberger's case--a cumulative normal psychometric function. But he found discrepancies (see his Tables XLVIII and XLIX). What happened was that he overlooked the fact that his range of stimulus levels did not push the response probabilities sufficiently close to 1.00 and .00. His computed expected limits in both ascending and descending directions were therefore biased and their average was in error. Fernberger's evaluating comments on the method of limits were therefore unjustified.

e. Preferred step size. Just as step size should be reasonably large for the up-and-down method, it appears that it should be similarly large when the method of limits is used to estimate μ. Large step sizes involve less bias of the one-way limit defined as the stimulus level midway between that of the last X response and that of the first Y response, involve less bias associated with initial testing level (a relation we may infer from Fernberger's data), and clearly require fewer trials per limit.

f. Estimate of μ when the psychometric function is not symmetrical. The average of an equal number of ascending limits and descending limits provides an unbiased estimate of the mean-median of the psychometric function when the latter is symmetrical. When there is assymmetry in this function, however, those series which start from the end where the function has the longer tail will result in an average limit which will depart more from the median of the psychometric function than those series which approach the median from the other direction. Hence the average of ascending and descending limits will deviate from the median of the psychometric function in the direction of the longer tail, just as m does in the case of the up-and-down method.

g. Continued observation by the same subject and the problem of his knowledge of the stimulus sequence. From the time of its earliest use, the method of limits posed a problem -- what to do about the subject's insight into or knowledge of the stimulus sequence from trial to trial. We have already discussed this problem with regard to the up-and-down method. As far as the method of limits is concerned, three ways of handling the problem have been advanced and adopted by different experimenters. (1) Give the subject full knowledge of the sequence. Wundt proposed this, arguing that only with knowledge of the sequence would the subject's attitude for observation be most favorable at the critical moment when the limit was reached. Guilford (1954) and Woodworth and Schlosberg (1954) support the use of the method in this form in spite of the recognition that successive judgments in the series will be interdependent. The view here is, in effect, that the object of the method of limits is to collect data under the condition of such interdependence. (2) Conceal or disguise the sequence of trials by the use of dummy trials between trials of the sequence, or by the use of several concurrent, interwoven series. These procedures are like those advanced for the up-and-down method of page 31 above. Fernberger (1913) did this in part, by interspersing trials by the method of limits with others by the method of constant stimuli. Others seem not to have followed his lead. On the one hand, our usual interest in both ascending and descending series makes the plan of using concurrent, interspersed series seem reasonable. On the other hand, the use of extreme testing levels to initiate series may introduce previous-trial stimulus context effects which could disturb our obtained limits. The strength of the latter argument would appear to make the concurrent series strategy much less suited to the method of limits (where extreme stimulus conditions are used) than to the up-and-down method (where most trials are conducted under non-extreme stimulus conditions). (3) Remove the sequence altogether by programming trials in random order. This routine was introduced by Kraepelin in 1891, and subsequently endorsed by Muller (1904), by Urban (1908) and by Titchener (1905) for at least some discrimination tasks. When data are collected in this way, the experimental session proceeds in the same manner as by the method of constant stimuli, but scoring is by the method of limits. The basic argument in favor of random ordering of the stimuli is that it prevents the development of any special "set" associated with the series, or as Smith recently puts it (1961) avoids "the effect of the subject's knowledge of the stimulating conditions on other than the relevant sensory basis." Random sequences must still be extended to extreme stimulus levels, however, if scoring by the method of limits is to be employed. Otherwise bias from restricted range will occur.

All in all, then, knowledge of the sequence is more difficult to deal with effectively in the method of limits than in the up-and-down method.

h. Statistical properties of the estimates of μ. In the case of m for the up-and-down method, we have seen that it is an approximate maximum likelihood estimate of μ. This property does not apply to the limit or the mean limit, but there is one interesting property which Urban (1908) observed for the "one-way" limit. Consider the various stimulus levels and the associated values of p_i, the probability of a Y response. When the expected distribution of limits is computed under the definition that the limit is the stimulus level where the first Y response occurs, the mode of that distribution cannot occur later than at the first stimulus level where p is greater than .50. Since it may occur before this level, however, and by an amount which depends on step size, this is not a very strong property of the limit.

i. Summary comments. The method of limits and the up-and-down method are similar in the sequential character of their testing programs and thus prove to have a number of comparable or parallel properties. They differ, however, in their procedures for estimating μ, and when this is the primary purpose of a study, our choice between using the "two-way" method of limits and the up-and-down method is clearly in favor of the latter on grounds of testing efficiency (See Anderson, et al, 1946).

18. FURTHER USES AND LIMITATIONS OF THE UP-AND-DOWN METHOD.

Thus far we have limited our discussion of the up-and-down method to its application in two-response, differential judgment situations. What we should like to do here is to comment briefly on the scope of the method.

a. Applications in stimulus threshold measurement and in scaling. The up-and-down method may reasonably be considered for application in any measurement situation where the probability of response function ranges from probability .00 (or close thereto) to probability 1.00 (or close thereto). The method will therefore serve well in single stimulus situations where the task is to determine stimulus or detection thresholds—the Bekesy and Oldfield problem. The method may also be applied in scaling work where the task is to locate limens between adjacent response categories. This particular case has been of interest to McDiarmid and is represented in some of the work to be described in Part II of this report.

Suppose we offer our subject the use of a number of ordered response categories--providing him with, say, four numbered response keys and requiring that he rate the loudness of each stimulus which he hears on a scale from 1 to 4. For this situation, the testing program is based on four-minus-one or three concurrent up-and-down series. For one of these series, each new trial moves up one stimulus step (to a more intense tone) following a response of "1", but moves down one step following a response of "2", "3", or "4". This series hunts for the limen between response categories "1" and "2". Similarly the second series moves up one stimulus level after a response of either "1" or "2", but down after "3" or "4". The third series moves down only after a response of "4" and moves up otherwise. The order of taking trials from the three series is random.

In our work, we have used this scaling procedure to evaluate the variability of a subject's differential judgments. He listened to pairs of stimuli and in judging the difference between them was allowed four response categories, which may be paraphrased as follows: (1) "certain the second tone was louder," (2) "thought the second one louder," (3) "thought the second one softer," (4) "certain the second was softer." As will be seen later, this technique provided very informative data relative to our stimulus context problem.

b. The location of percentile points other than the median of the psychometric function. In principle, the up-and-down method can be used to estimate any percentile point on a probability of response curve. Normality and a .00 to 1.00 probability range are assumed (Anon., 1944). It is recognized, however, that other sequential methods should be superior to the up-and-down method for the determination of high or low percentile points, methods which will concentrate trials in the vicinity of the desired points rather than near the median. Such methods include variants on the up-and-down method and other "staircase" methods, including the one-way method of limits (see Anderson, et al, 1946; Smith, 1961).

c. Problems where the probability-of-response curve does not fall to zero. The typical case where the psychologist is interested in estimating the 75th percentile point is not the case just discussed, but the case where the psychometric function ranges from probability 1.00 to .50 and levels off at the latter value. Such a situation arises, for example, when we would determine a stimulus threshold in a two-choice situation where random behavior results in 50% correct responses. Here, it is clear, the up-and-down method as such fails because there is no guarantee that the trial series will turn around at the lower end of the stimulus scale.

It might be thought, however, that modifications of the method would assure some success in keeping observations at the higher stimulus levels. Two modifications suggest themselves. One of these is to move up by large steps, say 2 to 4 times as large as the steps used in moving down. The other is to generate an up-and-down sequence on the basis of blocks of trials, where all trials in any block are run at the same stimulus level; e.g., move down one step if all trials in a 3-trial block are correct, but move up one step as soon as the first failure in a 3-trial block is observed. Examination of the expected trial distributions under these two modifications indicates that the second one should be the more effective of the two, provided that blocks of at least 3 trials are used. Analysis of the data would entail establishing a line of best fit to the response proportions at all stimulus levels, to provide the 75% point. This procedure, using only the programming feature of the up-and-down method, should be superior to the method of constant stimuli because it is sure to concentrate trials appropriately. (We note in passing that m computed for this "blocks up-and-down" sequence will not estimate any particular percentile point of the psychometric function, but may prove a useful statistic for comparing experimental conditions, as it did for Heinemann, 1961).

d. The context problem. Because the up-and-down method does not require many trials per measurement, it is particularly well suited to work on the effect of previous trials on a current judgment. The strategy is to use several concurrent series with the same subject, and to set up these series so that they differ as to standard, preceding trial condition, etc. Specific features of these testing procedures will be described in Part II.

19. SOME COMMENTS ON THE VARIABILITY OF s.

a. The estimated standard error of s. On the basis of the fact that s is an approximate maximum likelihood estimate of σ, the Princeton Research Group (Anon, 1944; Dixon and Mood, 1948) derived a suitable estimate of the standard error of s:

$$s_{s_{(\text{up-down})}} = \frac{s_{(\text{up-down})}H}{\sqrt{N_c}},$$

where H is a quantity which varies with step size and with the position of μ relative to the testing levels. H is almost everywhere larger than 1, and over the range of step sizes of greatest interest to us (i.e. between 1σ and 2σ), H has an average value of about 1.3 and never exceeds 1.4. If we are satisfied to know H to within 10%, we may take it to be 1.3 regardless of the position of μ relative to the testing levels. If we would be conservative, we may take the value 1.4.

It is of interest to recall that our estimate for the standard error of our traditional estimate of σ (i.e. $\sqrt{\Sigma x^2/(N-1)}$) is:

$$s_{s_{(\text{traditional})}} = \frac{s_{(\text{traditional})}}{\sqrt{2N}} = \frac{s_{(\text{traditional})}}{1.4\sqrt{N}}$$

Comparing this with the value for $s_{(\text{up-down})}$ when H = 1.3 or 1.4, we see that for a given value of s and a given value of N, the standard error of $s_{(\text{up-down})}$ is just about twice as large as the standard error of $s_{(\text{traditional})}$. Thus the variance of $s_{(\text{up-down})}$ is approximately four times as large as the variance of $s_{(\text{traditional})}$.

b. Implications for a test of homogeneity of values of s. From the above information we recognize that $s^2_{(\text{up-down})}$ cannot be distributed as χ^2

in the manner of $s^2_{(\text{traditional})}$. Hence tests of homogeneity of variance which we have for $s^2_{(\text{traditional})}$ are not directly applicable for $s^2_{(\text{up-down})}$. A test like Hartley's $F_{\max}$ test, however, would be a useful one to have, and this has led us to explore the possibility of adapting a Hartley-type test for evaluating the homogeneity of values of s from different up-and-down series.

The Hartley test (1950) is a "range test" on variance estimates. It is based on the fact that log s^2 (and hence also log s) is approximately normally distributed with variance equal to 2/(df-1). The distribution of the range of samples from a normal distribution is known for samples of any size, k, and so 95th percentile values of this range are known. What Hartley did was to define $F_{\max}$ as the largest of k values of s^2 divided by the smallest of the k values, and then derive the 95th percentile value of $F_{\max}$ from the 95th percentile value of the range as follows:

$$\left[\frac{\underline{\text{Range of k values}}}{\sigma}\right]_{.95} = \left[\frac{\log s^2_{\max} - \log s^2_{\min}}{\sqrt{\frac{2}{\text{df-1}}}}\right]_{.95} = \frac{\log\left(\frac{s^2_{\max}}{s^2_{\min}}\right)_{.95}}{\sqrt{\frac{2}{\text{df-1}}}}$$

where df represents the degrees of freedom in each estimate of s^2, and where $(s^2_{\max}/s^2_{\min})_{.95} = F_{\max_{.95}}$.

Now if we are willing (1) to presume that log $s_{(\text{up-down})}$ is about twice as variable as log $s_{(\text{traditional})}$, preserving the relative variabilities already observed for $s_{(\text{up-down})}$ and $s_{(\text{traditional})}$, and (2) to proceed as if log $s_{(\text{up-down})}$ were normally distributed, then we have

$$\left[\frac{\underline{\text{Range of k values}}}{\sigma}\right]_{.95} = \left[\frac{\log s^2_{\max_{UD}} - \log s^2_{\min_{UD}}}{2\sqrt{\frac{2}{\text{df-1}}}}\right]_{.95} = \frac{\log\left(\frac{s^2_{\max_{UD}}}{s^2_{\min_{UD}}}\right)_{.95}}{2\sqrt{\frac{2}{\text{df-1}}}}$$

$$= \frac{\frac{1}{2}\log\left(\frac{s^2_{\max_{UD}}}{s^2_{\min_{UD}}}\right)_{.95}}{\sqrt{\frac{2}{\text{df-1}}}} = \frac{\log\left(\frac{s_{\max_{UD}}}{s_{\min_{UD}}}\right)_{.95}}{\sqrt{\frac{2}{\text{df-1}}}}$$

This says that to obtain an approximate test of the homogeneity of several values of $s_{(up-down)}$, we should find the ratio of the largest to the smallest values of $s_{(up-down)}$--not the ratio of values of $s^2_{(up-down)}$--and use Hartley's tabled values of F_{max}. We find that this test involves a statistic which appears to be $\sqrt{F_{max}}$ rather than F_{max}. But when we remember that $s_{(up-down)}$ is itself basically a variance measure (see page 18 above), we see that the intended statistic is in reasonable form at that. In fact, it closely resembles an F_{max} based on variances of the trial distributions used for computing m and s.

Thus far we have a suggested testing routine for cases where step sizes are between 1σ and 2σ, but since the changes in H for small step sizes would only make the test more conservative, the test may prove useful for all cases with step sizes smaller than 2σ.

c. Evaluation of the proposed test. Of course what we really need is information on how well the test works in practice. Good evidence for this purpose should be available from the values of s obtained for concurrent series. Such series, taken over the same time period on the same subject, are homogeneous in their variance on a priori grounds. How does the test fare with such data?

Consider the 40 subject-sessions, each with three concurrent up and down series, which entered into our analysis in Table 9 above for experiments CS-1 and CS-2. Of these 40 sessions, 16 were conducted with steps between 1s and 2s in size, and 22 were conducted with steps smaller than 1s. The results of applying our test of homogeneity of values of s to these 38 sets of concurrent series are summarized in Table 13. Critical values of F_{max} were found by interpolation in available tables (Walker and Lev, 1953) for an N of 14 for experiment CS-1 and an N of 19 for CS-2. The proportions of "non-homogeneous" sets of s values appear to be reasonably close to the intended .05 and .01 rejection levels. A further check of the distribution of log (s_{max}/s_{min}) for these sets of data against the distribution of the range for samples of three from the normal distribution (McKay and Pearson, 1933) indicates that the rejection levels should hold fairly well. We thus have some empirical support for the use of the (s_{max}/s_{min}) criterion of homogeneity.

Of course we shall never be concerned with testing the homogeneity of concurrent series. Rather our interest will be in testing homogeneity from day to day with the same subject or from replication to replication in

Table 13

EMPIRICAL CHECK ON REJECTION RATE WHEN s_{max}/s_{min}
IS USED TO EVALUATE HOMOGENEITY OF VARIANCE (for K=3).

Range of step sizes	No. of sessions	No. of concurrent series per session (i.e. no. of s values)	Sessions where (s_{max}/s_{min}) exceeded $F_{max\,.95}$		Sessions where (s_{max}/s_{min}) exceeded $F_{max\,.99}$	
			Number	Proportion	Number	Proportion
1s - 2s	16	3	2	.125	0	.00
below 1s	22	3	1	.045	0	.00
				(.05 expected)		(.01 expected)

(Data from experiments CS-1 and CS-2)

one-trial-per-subject experiments. Such a concern arose in experiment T-1 above (pages 35-38) and we there used the ratio s^2_{max}/s^2_{min}. We note here that use of the present s_{max}/s_{min} criterion leads to the acceptance of the hypothesis of homogeneity for all three sets of data which were discussed there.

20. SUMMARY AND EVALUATION.

This part of our report has been concerned with a description and evaluation of the up-and-down method. We have seen it to be a sequential method, applicable to the study of absolute or stimulus thresholds, differential discrimination, context and scaling problems.

Early sections of our discussion dealt with the characteristics and properties of the method under the model that the probability of response curve or psychometric function is a cumulative normal distribution. Estimates for μ and σ were reviewed and the effect of using different sizes of stimulus steps was discussed. It was indicated that the preferred stimulus step for much psychological research will be between 1σ and 2σ in size.

Subsequently, two testing arrangements were discussed: one in which each trial on the up-and-down series is conducted using a different subject (the one-trial-per-subject procedure), and the other in which all observations are made on the same subject and the up-and-down sequence is concealed through the use of several concurrent series (the concurrent series procedure).

For the one-trial-per-subject procedure, variability of the up-and-down series is a function of within subject variability and between subject variability, as summarized in Table 14. Successive trials are clearly independent and the normality aspect of the model for the up-and-down method would appear to be satisfied with a suitable choice of stimulus scale. As indicated at the bottom of Table 14, the estimate of the standard error of m as provided by the up-and-down method gives indications of being conservatively large.

When the up-and-down method is used to study the discrimination of a single subject, the model cannot be fully satisfied if interdependencies exist between successive trials. Although the use of concurrent series serves to eliminate interdependencies arising from insight into the character of the up-and-down series, there is no way of avoiding dependencies associated with drift of the PSE or with the effect of the stimuli for each

Table 14

VARIANCE COMPONENTS AND VARIABILITY MEASURES IN ONE-TRIAL-PER-SUBJECT TESTING.

	Variability measures			
A. Sources of variance:	s: for a single one-trial-per-subject series	s_m: as computed from $\bar{s}$	$s_{m_{obs(SG)}}$: observed variability of m for replications with the Same Group of subjects	$s_{m_{obs(IG)}}$: observed variability of m for replications with Independent Groups of subjects
	(* indicates that source contributes to var. meas.)			
Variance of the psychometric function for the individual subject, at any given time.	*	*	*	*
Variance of μ for the individual subject from time to time (i.e. drift in the location of the probability of response curve): the basis of sampling variance assoc. with time of testing each subject.	*	*	*	*
Variance in μ from subject to subject in a given sample of subjects.	*	*	*	*
Variance in μ between independent groups of subjects, i.e. sampling variance with regard to subjects.				*
B. Experimental Results:				
For 3 experiments		s_m >	$s_{m_{obs(SG)}}$	
For 1 experiment		s_m	>	$s_{m_{obs(IG)}}$

trial upon the judgment on the trial to follow. Of course all psychophysical measurement must either live with these dependencies or quantify them in some way. With the up-and-down method we have a procedure which should let us examine and quantify these dependencies in an efficient manner.

For the use of concurrent series with the same subject, the variabilities with which we have to deal are those summarized in Table 15. This table includes reference to observed variability between concurrent series ($s_{\bar{m}_{obs(con)}}$), which represents a "split-halves" kind of reliability, and to observed session-to-session variability ($s_{\bar{m}_{obs(sess)}}$), which is a test-retest reliability measure. We find, as in the other areas of psychological research, that split-halves reliability often exceeds test-retest reliability.

Evidence has been presented to show that for the determination of a PSE, the up-and-down method has advantages over both the method of constant stimuli and the method of limits. It provides measures which may be described as more valid than measures by the method of constant stimuli, where the obtained PSE is constrained to be near the POE. The up-and-down method also provides PSEs which are potentially less biased then PSEs obtained by the method of limits. Because the up-and-down method hunts for the PSE and concentrates trials near the PSE, satisfactory measurements can frequently be made with this method on the basis of a very small number of trials.

Table 15

VARIANCE COMPONENTS AND VARIABILITY MEASURES
WHEN CONCURRENT SERIES ARE RUN WITH THE SAME SUBJECT.

	Variability measures			
	s: for a single up-and-down series	s_m and $s_{\bar m}$, as computed from $\bar s$	$s_{m_{obs(con)}}$: observed variability of m for concurrent series	$s_{\bar m_{obs(sess)}}$: observed variability of $\bar m$ for sessions.
A. <u>Sources of variance</u>:	(* indicates	that source	contributes	to var. meas.)
Variance of the probability of response curve or psychometric function of the subject at any given time.	*	*	*	*
Variance of μ for that subject from time to time in the same test session (i.e. trial-to-trial variability or drift in the location of the probability of response curve).	*	*	*	*
Covariance between concurrent series.			*	
Variance in μ for the subject from session to session (i.e. long-term changes in the location of the probability of response curve).				*
B. <u>Direction of Experimental Results:</u>				
Positive covariance between series, when step size is small. This makes. . .		s_m >	$s_{m_{obs.}}$	
Large session changes for some subjects. This makes. . .		$s_{\bar m}$	<	$s_{\bar m_{obs.}}$

References

1. Anderson, T. W., McCarthy, P. J., and Tukey, J. W. *"Staircase" methods of sensitivity testing*. NAVORD Rept. 65-46. (Statist. Res. Group, Princeton) 21 March, 1946. pp. 134.

2. Anon. *Statistical analysis for a new procedure in sensitivity experiments*. App. Math. Panel Rept. No 101.1R, (Statist. Res. Group, Princeton: No. 40), July 1944. pp. 58.

3. Bekesy, G. V. A new audiometer. *Acta*. *Oto-laryngol*., 1947, *35*, 411-422.

4. Blough, D. S. Dark adaptation in the pigeon. *J*. *comp*. *physiol*. *Psychol*., 1956, *49*, 425-430.

5. _______. Spectral sensitivity in the pigeon. *J*. *opt*. *Soc*. *Amer*., 1957, *47*, 827-833.

6. Brown, J. and Cane, V. R. An analysis of the limiting method. *Brit*. *J*. *Statist*. *Psychol*., 1959, *12*, 119-126.

7. Brownlee, K. A., Hodges, J. L., and Rosenblatt, M. The up-and-down method with small samples. *J*. *Amer*. *Statist*. *Assn*., 1953, *48*, 262-277.

8. Christensen, K. Isohedonic contours in the sucrose-sodium chloride area of gustatory stimulation. *J*. *comp*. *physiol*. *Psychol*., 1962, *55*, 337-341.

9. Cornsweet, T. N. The staircase-method in psychophysics. *Amer*. *J*. *Psychol*., 1962, *75*, 485-491.

10. Day, W. F. *Stimulus interval as a determiner of serial patterns in threshold responses*. Master's Thesis, Univ. Virginia, 1951.

11. Dixon, W. J. and Massey, F. J. *Introduction to statistical analysis*. Second Ed. New York: McGraw Hill, 1957. (Earlier edition, 1951).

12. Dixon, W. J. and Mood, A. M. A method for obtaining and analyzing sensitivity data. *J*. *Amer*. *Statist*. *Assn*., 1948, *43*, 109- .

13. Evans, W. O. *A titration schedule on a treadmill*. U. S. Army Med. Res. Lab., Rept. No. 525. 21 Dec 1961. pp. 7.

14. Fernberger, S. W. On the relation of the methods of just perceptible differences and constant stimuli, *Psychol*. *Monogr*., 1913, *14*, No. 61, pp. 81.

15. Gourevitch, G., Hack, M. H., and Hawkins, J. E. Auditory thresholds in the rat measured by an operant technique. *Science*, 1960, *131*, 1046-1047.

16. Guilford, J. P. *Psychometric methods*. Second Edition. New York: McGraw Hill, 1954. pp. ix + 597.

17. ________. System in the relationship of affective value to frequency and intensity of auditory stimuli. *Amer. J. Psychol.*, 1954, *67*, 691-695.

18. Harris, J. D. Discrimination of pitch: suggestions toward method and procedure. *Amer. J. Psychol.*, 1948, *61*, 309-322.

19. Hartley, H. O. The maximum F-ratio as a short cut test for heterogeneity of variance. *Biometrika*, 1950, *37*, 308-312.

20. Heinemann, E. G. The relation of apparent brightness to the threshold for differences in luminance. *J. exper. Psychol.*, 1961, *61*, 389-399.

21. Kappauf, W. E., Burright, R. G., and DeMarco, W. Sucrose-quinine mixtures which are isohedonic for the rat. *J. comp. physiol. Psychol.*, 1963, *56*, 138-143.

22. Kappauf, W. E. and Payne, M. C. *Intraserial correlations in brightness matching responses*. Memo. Rept. H-3. Contract AF 33(038)-25726, Univ. Illinois, 1954, pp. 14.

23. _______. *Intraserial correlations in brightness matching responses: II*. Memo. Rept. H-7. Contract AF 33(038)-25726. Univ. Illinois. 1955, pp. 14.

24. Koester, T. The time error and sensitivity in pitch and loudness discrimination as a function of time interval and stimulus level. *Arch. Psychol.*, 1944-45, *41*, No. 297.

25. Koester, T. and Shoenfeld, W. N. The effect of context upon judgments of pitch differences. *J. exper. Psychol.*, 1946, *36*, 417-430.

26. Koh, S. D., and Teitelbaum, P. Absolute behavioral taste thresholds in the rat. *J. comp. physiol. Psychol.*, 1961, *54*, 223-229.

27. Kohler, W. Zur Theories des Sukzessivergleichs und der Zeitfehler. *Psychol. Forsch.*, 1923, *4*, 115-175.

28. Lange, K. A recording sphygmotonograph: a machine for the continuous recording of systolic and diastolic arterial pressure in man. *Annals of Internal Medicine*, 1943, *18*, 367-383.

29. Loeb, M., and Dickson, C. Factors influencing the practice effect for auditory thresholds. *J. acoust. Soc. Amer.*, 1961, *33*, 917-921.

30. McCarthy, P. J. A class of methods for estimating reaction to stimuli of varying severity. *J. educ. Psychol.*, 1949, *40*, 143-156.

31. McKay, A. T. and Pearson, E. S. A note on the distribution of range in samples of n. *Biometrika*, 1933, *25*, 415-420.

32. McDiarmid, C. G. *Context effects in differential loudness judgments*. Ph.D. Thesis, Univ. Illinois. 1962(a), pp. 117.

33. ________. *Context effects in differential loudness judgments*. Report under Contract DA-49-007-MD-877. Univ. Illinois. 1962(b), pp. 67.

34. Michaels, W. C., and Helson, H. A. A quantitative theory of time-order effects. *Amer. J. Psychol.*, 1954, *67*, 327-334.

35. Muller, G. E. *Die Gesichtspunkte und die Tatsachen der psychophysichen Methodik*. Wiesbaden: J. F. Bergmann, 1904.

36. Needham, J. G. The time error in comparison judgments. *Psychol. Bull.*, 1934, *31*, 229-243.

37. ________. The time error as a function of continued experimentation. *Amer. J. Psychol.*, 1934, *46*, 558-567.

38. ________. The effect of the time interval upon the time error at different intensive levels. *J. exper. Psychol.*, 1935, *18*, 530-543.

39. Oldfield, R. C. Continuous recording of sensory thresholds and other psycho-physical variables. *Nature*, 1949, *164*, 581.

40. Postman, L. The time error in auditory perception. *Amer. J. Psychol.*, 1946, *59*, 193-219.

41. Pratt, C. C. The time error in psychophysical judgments. *Amer. J. Psychol.*, 1933, *45*, 292-297.

42. Smith, J. E. K. Stimulus programming in psychophysics. *Psychometrika*, 1961, *26*, 27-33.

43. Stevens, S. S. The direct estimation of sensory magnitudes. *Amer. J. Psychol.*, 1956, 69, 1-25.

44. ________. On the psychophysical law. *Psychol. Rev.* 1957, *64*, 153-181.

45. Stott, L. H. Time-order errors in the discrimination of tonal durations. *J. exper. Psychol.* 1935, *18*, 741-766.

46. Titchener, E. B. *Experimental Psychology*. New York: Macmillan, 1905. Vol II, parts 1 and 2.

47. Urban, F. M. On the method of just perceptible differences. *Psychol. Rev.*, 1907, *14*, 244-253.

48. ______. *The application of statistical methods to the problems of psychophysics*. Phila: the Psychological Clinic Press, 1908. pp. ix + 221.

49. Verplanck, W. S., Collier, G. H. and Cotton, J. W. Non-independence of successive responses in measurements of the visual threshold. *J. exper. Psychol.*, 1952, *44*, 273-282.

50. Votaw, D. F. *The effect of non-normality on "Staircase" methods of sensitivity testing*. (Statist. Res. Group, Princeton). 1 May 1948. pp. 39.

51. Walker, H. M., and Lev, J. *Statistical Inference*. New York: Holt, 1953, pp. xi + 510.

52. Woodrow, H. Weight discrimination with a varying standard. *Amer*. *J*. *Psychol*., 1933, *45*, 391-416.

53. Woodrow, H., and Stott, L. H. The effect of practice on positive time order errors. *J*. *exper*. *Psychol*., 1936, *19*, 694-705.

54. Woodworth, R. S., and Schlosberg, H. *Experimental Psychology*, Rev. Ed. New York: Holt, 1954. pp. xi + 948.

55. Young, P. T. Isohedonic contour maps. *Psychol*. *Rep*., 1960, *7*, 478.

56. Young, P. T., and Schulte, R. H. Isohedonic contours and tongue activity in three gustatory areas of the rat. *J*. *comp*. *physiol*. *Psychol*, (in press).